Assessing the Validity of Multiple-Choice Questions, Using them to Undertake Comparative Analysis on Student Cohort Performance, and Evaluating the Methodologies Used

Jacob Rhys Marchant
School of Physical Sciences, Department of Chemistry

December 2020

THE UNIVERSITY of ADELAIDE

Contents

Abstract

Multiple-choice questions (MCQs) have been used in all four first-year Chemistry courses at The University of Adelaide for many years to assess the students both during the semester and within the final examinations. Ensuring that the assessment tasks and the items used within them are a valid method of determining student competency is an important part of reviewing the results of the assessments. Throughout this research all the MCQ assessment tasks and items used in first-year Chemistry courses between 2012-2015 were reviewed and analysed using Classical Test Theory (CTT) and Rasch Analysis. Out of 261 unique items that were utilised in MCQ assessment tasks across the four first-year Chemistry courses over the 4-year period analysed most of the items were found to be performing appropriately. However, 12 of them were found to be consistently problematic by CTT (4 major issues, 4 potentially major issues, and 4 minor issues), and 83 items were seen to be consistently problematic using Rasch analysis (41 major issues, 9 potentially major issues, and 33 minor issues), showing a large difference in the items that are identified by each methodology. Due to the fact that these problematic items are spread over a large number of assessment tasks that took place during different years none of the individual assessment tasks contained enough problematic items that it impacted upon the validity of those assessment tasks.

After excluding these problematic items when analysing for differences in how male and female students perform 27 unique items were identified to contain gender bias using CTT (18 male, 4 female, 5 alternating) out of 249 unique MCQ items while only 14 unique items were identified to contain gender bias by Rasch analysis (7 male, 6 female, 1 alternating) out of 178 unique items considered.

The performance of student cohorts over different time intervals was also explored within this research. This involved the comparison of students' results between the lecture tests undertaken during the semester to their results within the MCQ section of the final examination. It also involved comparing student cohorts from the same course over different years to determine if student competency changes significantly between. On average students performed to a higher level on MCQ assessment tasks later in the semester years (i.e. a 'test-retest' strategy is beneficial for students). There were no discernible trends identified when comparing between yearly student cohorts within the same course, thus indicating that student competency did not change significantly within the period studied (2012-2015). Comparing between Biology and Chemistry courses was also attempted within this research; however, due to complications in linking the results of the assessment tasks it was determined that this was not possible using the available data.

As the entirety of this research was based around two different methodologies that can evaluate MCQ assessment tasks and items in different ways (Classical Test Theory and Rasch Analysis) it gave the opportunity to compare the two methodologies. Classical Test Theory was found to be a highly approachable methodology that provided a high-level overview of an assessment task and its items, making it a suitable methodology for analysing low stakes assessment tasks to improve upon any large issues within those tasks. Rasch Analysis could be used to analyse the performance of students as well as the assessment tasks and items, allowing for greater versatility in how the analysis can be applied and the information that it can provide. This makes Rasch Analysis a more appropriate methodology for high stakes assessment tasks where it is important that the most accurate reflection of the students' abilities and any issues present within the task need to be identified and addressed.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Acknowledgements

I would like to thank my supervisors Simon Pyke and Natalie Williamson for their continued assistance and guidance throughout this research. Without their help none of this would have been possible and I sincerely thank them for that.

My family and friends have also been of great assistance to me throughout this entire process through their support and keeping me grounded. They have always encouraged me to strive for my best and believed in me even when I doubted myself.

Finally, none of this research would have been possible without the students who undertook the assessment tasks that are the foundation of this entire research. Therefore, all those students deserve the biggest thank you as they were unknowingly the biggest contributors to this work.

Chapter 1. Introduction

1.1 The Purpose of Assessment

1.1.1 Why Assess?

The purpose of an assessment is influenced by the objectives of the assessment, as an assessment can be used to help students learn, determine students' progress, or to assist educators in making decisions in regard to how they approach educating the students.[1-3] Assessment helps students learn as it informs them of their own progress toward learning objectives, providing them with the opportunity to reflect and learn based on their performance. The determination of student progress may either be used as an evaluation of the teaching process (determining how well students have learnt the material being covered), or it could be used as a measure of the students' competency within the content being assessed, which may then be used to assign a grade to a student. Assessment assisting educators ties in with the idea of an assessment measuring the effectiveness of the teaching pedagogy, as it may help to inform educators of areas that the students require more support in, or it can be used to determine what teaching pedagogy the students are responding to. Both possibilities can be used to better improve how the educators teach the students. None of these options are mutually exclusive, as an assessment may be used to determine student learning as well as the effectiveness of the teaching pedagogy. The objective of the assessment will influence what the main purpose of that assessment is, and it may also influence when the most appropriate time to assess the students is.[4] This is the reason that the purpose of an assessment needs to be considered when implementing an assessment, as its purpose will help to shape the construction of the assessment.

1.1.2 What is Assessment?

There are two distinct types of assessment, "assessment of learning" and "assessment for learning".[3] Assessment of learning is primarily used to measure what the students have learned within a specific area, which can then be used to rank students, provide feedback, or assign grades. Assessments that affect the final grade or rank of the students are termed "summative assessments" and typically occur after teaching has occurred to determine if learning has occurred, making them assessments of learning.[3] Assessments for learning are intended to help identify student progress in their learning and how they need to progress to show competency, thus helping both students and teachers. These sorts of assessments typically give the students an opportunity to practice questions and obtain feedback, and thus are generally represented as "formative assessments".[3,5,6] It is important to utilise both formative and summative assessments within a course (in this context a course refers a semester's worth of content), as formative assessments should be used to inform the students on which areas that they can improve upon before they undertake a summative assessment. Usually summative assessments do not provide as much feedback to the students as formative assessments do, as it is common that some aspects of summative assessments are reused from year to year and thus making the answers public or giving detailed feedback is a concern for the validity of future assessments. However, the students benefit from some amount of feedback from any assessment to help them improve their performance.[3,5-8] Both types of assessments provide information to the assessors about the students' knowledge of the course content and areas of concern, which can be used to inform teaching practices, as well as some amount of information about the effectiveness of the teaching pedagogy.

1.1.3 Impacts of Assessment

It is well understood that assessment drives student learning.[9-22] This occurs primarily as a result of three major influences that assessments have on student learning. Firstly, the timing and the content of the assessment determines at what point in the course the students will attempt to reinforce and/or relearn the content that is being assessed. Secondly, the feedback provided to the students informs them of what they need to work on to improve their future results. Thirdly, this influence arises from the assessors' reaction to the assessment results, as based on the results of the assessment, the educators may change how, what, or when the students are taught different aspects of the course and the focus of those teachings. The combination of these influences means that assessment is a strong driving force for what the students spend their time learning, and when they learn it.

Assessment evolves with education level, as in primary and early secondary education levels the actual rank or measure of the students' abilities is less important than them meeting a predefined level of expectation (i.e. achieving a minimum level of performance). However, in higher secondary and tertiary levels of education the rank and the measure that the student is awarded can significantly affect the future pathways available to a student, which in turn can considerably influence student outcomes upon finishing their education. These outcomes can be related to decisions that students make in regards to future education, careers, and employment. That is not to say that what is possible for a student is dictated entirely by their results in assessments, but rather assessments can shape what a student and others believe they are capable of. This can have lasting effects on that student's life, highlighting the need to ensure that assessment outcomes are a fair and accurate representation of ability.

Assessment tasks need to fit into the syllabus of the courses, as the students need to be studying the course content not the assessment itself to ensure that the results are reflective of their ability within the course and not just within the assessment.[9,23] Some students will spend more time studying the best way to answer specific questions that they believe will be in the assessment rather than learning the course content, which is mainly the result of two factors. Firstly, students seeking to maximise their marks with the smallest possible effort (strategic learners) will engage in this style of learning because it is seen as the most efficient way to earn marks within an assessment.[24-27] Secondly, if the assessment is viewed by the students as an unfair method for gauging their ability within the course, then the students will shift away from learning the course content and attempt to learn the assessment as the two are not seen to be mutually inclusive.[9,10,14,28] While it is hard to shift the strategic learners away from this style of learning, if the assessment is typically seen as a fair way to assess the syllabus, then most students believe that learning the course content gives them their best chance of success.[18-20,27]

It is important to consider the emotional impact that assessments can have on the students and how that may be reflected in the performance of a student. Anxiety and stress can be felt by students because of assessment, particularly if the assessment has a large influence on the student's final grades or their future opportunities. These feelings can negatively impact the students' performance to such an extent that their performance in an assessment is not reflective of their actual ability.[29-31] In many cases students will be anxious and stressed before they undertake the assessment, but once the assessment begins they tend to calm down and are able focus on the assessment.[32] In other cases, these emotions are accounted for within the assessment criteria, as being able to compose themselves and perform may be an important aspect of an assessment task. For example, when

performing in front of an audience it is common that students will be nervous; however, performing well despite those nerves is an important and assessable part of those forms of assessment. Despite this, it is important that these emotions never impact the student in such a way that the construct validity of the assessment (the degree to which the assessment measures what it aims to be measuring; e.g. the student's understanding in a specific area) is negatively impacted. Thus, it is an important consideration when generating assessments how a student will react emotionally to that assessment, and if that can influence measures of the students' performance within the assessment.

Stemming from self-confidence impacting the results of an assessment task is that students may self-select based on their results in assessment tasks. Self-selection is when the students select courses and topics that they believe fit with their own goals, values, and ability.[33] This idea becomes more relevant at higher levels of education where students have more influence over the courses that they are studying, and thus have the ability to customise their enrolment to suit them. Typically, students will choose courses based around some amount of self-selection, which can be due to their career choice, their results in previous assessments, or a variety of other factors. The results of previous assessments can influence student confidence both positively and negatively, and thus may result in students either continuing a course or selecting something different because of the results obtained, and thus it is important that assessment results are an accurate reflection of student ability.

The impact of assessments on students is both expected and necessary to obtain measures that can measure the performance of the students. However, while these impacts may be necessary, that does not mean that they should not be considered as an influential factor within assessment. The influence that assessments can have over student decisions is both substantial and impactful in regards to both the present and the future of any student.[33] This is why it is imperative that assessors deeply consider the assessment tasks that they are using to ensure that the decisions that the students are making based on the results of those tasks are being made based on a fair and accurate representation of their ability.

1.1.4 Assessment Formats

An assigned rank or grade is often used as a measure of that student's competency within the course or topic that is being assessed. The ranking or grade could be related to any aspect of the students' education, and thus assessments need to be flexible enough to be able to measure the student's ability based on any knowledge, skills, or competencies that are expected within the course. The assessment must also be an effective way to test these traits to ensure that any impact on the students' results is due to the students' ability rather than an unrelated factor, which often will require several different assessment formats to ensure all aspects are covered to some extent. There are several assessment formats that attempt to measure these traits, with some being able to target some traits better than others. Very broadly there are five types of assessment formats that involve an important written aspect: multiple-choice questions (MCQs), short-answer questions, long-answer questions, written response, and practicals. Each of these formats assesses the students in different ways, and how each of them is used is dependent on the course.

Multiple-choice questions (MCQs) are most commonly presented to students using the single-best response format, where the students choose the option that best answers the question asked of them out of a list of possible responses. There are a multitude of other MCQ formats that can be used in assessment (Section 1.4.3), which should be chosen based on the requirements of the assessment.[34] MCQs are seen as an effective way to assess large amounts of students on a broad

range of topics in a short period of time.[35-38] This is because the items (the information presented to the student including the question, the options, and any additional information) can be answered relatively quickly by the students and thus they are able to answer a large number of items in a single assessment, which can then be quickly marked by the assessors.

Short-answer questions are the most common form of constructed response, as they provide the students an opportunity to explain their thought process while being fast enough to allow for the coverage of several different topics. They typically require the students to write a short paragraph or display their working on a problem that allows them to exhibit their knowledge and understanding of what is being asked of them. This provides a better overview of the student's ability on a specific topic or concept, and thus is highly effective at identifying student misconceptions and misunderstandings.[39,40]

Long-answer questions or essay style questions require the students to answer a question in a large amount of detail, and as such these questions tend to be weighted higher as the student is expected to show a deeper understanding when answering. This format is most commonly used with the intention of assessing a student's deeper understanding of a topic or concept, as the students can be asked to evaluate and generate their own ideas to answer the question.[39] These types of questions also provide the students with an opportunity to demonstrate the depth of their ability that most other formats cannot replicate.

A written response assessment could be a more detailed version of a long-answer question, several short-answer questions, a self-directed investigation, a creative writing piece, or some other form of writing. This type of assessment task is undertaken outside of test conditions, which means that more detailed work is generally expected from the students, and students are given an excess of time to complete any tasks given in this format. As a result of this, it is expected that the students are able to fulfil the criteria of the assessment task and present in it in an appropriate manner within whatever time frame is given.

Practical assessments allow students to clearly demonstrate their ability to perform required outcomes of the course. However, practical assessments are not a viable form of assessment within every course. This is because in some courses, assessing students on a practical aspect is either irrelevant or difficult, depending upon the course content. However, in courses where students can be assessed using a practical assessment, this aspect of the course usually represents a competency that is required to complete the course. Practical assessments involve placing students in situations that are related to the type of work they would be expected to undertake in employment relating to the course. For example, student-teacher placements, lab experiments, mock trials, and presentations are all examples of practical assessments. These assessments are important in ensuring that students gain experience in and can complete the sorts of tasks that would be expected of them in employment.

1.2 Assessment Format Considerations

1.2.1 Why Different Formats?

It is important to select a variety of assessment formats that together assess the students' completion of the educational objectives of the course, as each format differs in how it assesses the students and its logistical requirements. This means that the requirements of both the course and

the assessment need to be considered when choosing an assessment format, which is tied into the idea of face validity of an assessment. There is no point in the students completing a ten-page written response if the assessors do not have the time to mark them. Similarly, there is no reason for the students to undertake a MCQ assessment if the objective of the course is purely about the student obtaining practical experience in their field. Therefore it is important to select a format that aligns with the objectives of the course and fits within any restrictions placed on the assessors to ensure the validity of the task.

1.2.2 Multiple-Choice Questions

Any MCQ is made up of four components: the stem, which is the question that is being asked of the students; the key, which is any option(s) that are considered correct; the distractors, the option(s) that are considered to be incorrect; and any supplementary information that is provided within the item. The construction of each individual component may vary between or even within an assessment task, which can be used to create versatile and unique assessments. MCQs are the most efficient and economical of all of the assessment formats due to the speed at which they can be undertaken by the students and marked by the assessors.[34,38,41,42] However, most of the time spent on a MCQ assessment by assessors should be before the assessment takes place when the items used are generated.[43,44] The difficulty of writing new MCQs is often underestimated, and generating new and effective MCQs takes a substantial amount of time.[38,45,46] This is because it is important to ensure that there are no construction issues within the item (e.g. double negatives, answer cueing, unfocused stem, etc.), and that the item assesses the students on the desired content. As a result of this, each MCQ should be extensively reviewed after it is written to ensure that any potential issues are minimised.[38,43,45,47,48] The time spent by assessors on writing MCQs can be mitigated through the use of an item bank, which may contain a large amount of items that are rotated between assessment tasks or only enough items for one assessment task. This negates the need for new items each time an assessment takes place; however, using an item bank is not sensible for every course, especially if the items are made public after the assessment has taken place. An item bank does not preclude the need to evaluate the items both before and after they are used to ensure that they are relevant to the course and performing as expected. However, items banks are useful because, assuming the performance of items does not change between years, a large majority of the items used within a course do not need to be generated every time a MCQ assessment takes place, but rather only when new items are required to assess a new aspect of the course or when an old item needs to be replaced.

Due to the speed at which students can respond to a MCQ it means that a broad range of content can be assessed in a short amount of time. This means that within one assessment the students can be assessed on at least one concept from every aspect of the course, providing a rough indication of the student's ability across the entire course. However, the MCQ format does not provide much information about the students' reasoning for why they selected an option.[39,49] That being said, the items can still provide an insight into misconceptions that are held by the students, as often the incorrect options (the distractors) represent answers that would be obtained through misconceptions or flawed logic. If misconceptions represent different developmental states within the student's learning, through the strategic use of misconceptions as distractors it may be possible to gain a better understanding of where the students lie in their learning.[50-52] However, it can be hard to justify whether an option was chosen due to student guessing or whether the student genuinely thought they were selecting the correct option. This is because without having any sort of insight into the student's thought process the reason that they chose an option is almost always entirely conjecture.

MCQs are objectively marked, which means that if the item is properly constructed there is only one correct answer or a clear single-best answer within the options presented to the students. Depending on the type of MCQ used there is the potential for students to be awarded partial marks, but most commonly students will receive full marks for the correct answer and no marks if any other option was selected.[53,54] As MCQs are objectively marked, it means that marking the assessments can be automated, which allows for any MCQ assessment to be promptly marked. Additionally, any MCQs that are undertaken online can be marked immediately providing feedback to the student directly after they have undertaken the assessment, allowing the students to immediately identify any misconceptions they may have. Objective marking makes the MCQ format very appealing to assessors that have large cohorts of students, as they simply do not have the time required to mark hundreds of assessments every time the students are assessed.[55-57]

The lack of any information gained about the students' reasoning in conjunction with the use of an item bank can make it difficult to give students effective feedback from a MCQ assessment, as if the assessors are unsure where the students went wrong, they cannot correct the students' misconceptions. There is also a concern that MCQs do not award students for incomplete knowledge, as most commonly students are awarded either a 1 or a 0 on each item. Both of these factors means that generally students get little opportunity to display their knowledge, and they potentially receive no effective feedback, which is important to help them improve their understanding.[39,58,59] This means that from the students' perspective, MCQ assessments might not be ideal for them to display the extent of their knowledge, or to gain specific insight on where they need to improve. However, assessors will find MCQ assessments useful in almost every single course, as they can be used to assess the students' knowledge with minimum effort from the assessors once the items have been generated.

1.2.3 Constructed Response

Constructed response tasks include any format in which the students are required to generate their own answers by writing them to complete the assessment. As such, constructed response encompasses short-answer questions, essay/long-answer questions, and written response. Because the students need to generate the answer in their own words, these formats provide an insight into the student's thought process. Thus, constructed response can help to identify how deep a student's understanding of the content is and common misconceptions within the course.[39] Constructed response style questions would be expected to more accurately determine the student's level of understanding than MCQs, as more detail is required from the students when answering the question. Often short-answer questions are seen as the least probing while written responses are the most probing, as the longer the student has to construct a response the more detail that can be expected of them.[60] However, regardless of the level of understanding that they assess, all of the answers may differ significantly from student to student, and thus every constructed response format requires the assessment to be marked by hand. This is both time consuming and can be subjective, which means that particularly in the case when more than one assessor marks the work, it is important to have a well-defined performance standard or rubric. A well-defined rubric will not only ensure that all of the assessors are aware of what is expected of the students, but in some scenarios, it can also be used to inform the students what is expected of them. As a consequence, constructed response questions are more time consuming for the students to undertake per item asked of them and more time consuming for the assessors to mark than the MCQ format, which means that typically less content can be covered within a constructed response assessment than a

MCQ one. However, they are able to probe the student's thought process, and thus provide more information about the students' ability level per item.[61]

There is a concern within both long and short answer questions (less so with written response) that students are more likely to discuss ideas and concepts that they are familiar with and confident in.[40] This means that some students will provide a large amount of detail on one item, and then provide very little detail on another because they were unsure of the answer and more willing to spend time perfecting one answer than attempting to explain a concept that they are unsure of.[40,62] As such they will obtain full marks on an item in which they provided more detail than necessary, and less marks on another item that they were less confident in and spent less time on as a result. A student's lack of confidence does not mean that they would give an incorrect answer, and in reality, often students will receive more marks when they take 'educated guesses' than when they only provide answers they are confident in. There is also a tendency for students who are unsure of how to answer an item to rephrase the information given to them in the item, either in the hope that it will give them some marks or it will help them remember the answer. In comparison to this, there are different expectations placed on students when they undertake written response assessments. The lack of a time limit in which they need to answer the item (apart from the due date) means that there is an expectation that the students' answers are concise and only include relevant information.

There is a larger range of questions that are possible using constructed response compared to MCQs, and the level of detail required within the answer should be reflected in the format of the constructed response being used. If the students are expected to be able to answer the question in a small paragraph then this should be a short-answer question, but if more detail is required then it becomes more appropriate to assess using either an essay question or within a written response. It is also possible to assess different aspects of the students' knowledge and understanding using constructed response as opposed to other formats. For example, a MCQ requires a definitive answer to a question posed to the students, but within any constructed response format the students can be marked based on the process that they undertake rather than the final answer that they obtain. This theoretically allows for the allocation of partial marks on a specific item, as the assessors can identify where the student has made a mistake in their answer. If the mistake represents a small error (e.g. using the wrong numbers in an equation), but the majority of the item is answered correctly then the students will likely still receive the majority of the marks allocated to that item, as they still display an understanding of the relevant concepts.

1.2.4 Practical Assessments

Practical assessments are an extremely important aspect of some courses, and not utilised in other courses.[63,64] This is because the use of practical assessments is dependent on the course being undertaken, as only some courses require students to have experience in relevant tasks. Courses that do require students to show a level of competency in a task should have some form of practical assessment, but for any other course that may not require this, practical assessments may not align with assessment objectives and thus will not be utilised.

A practical assessment should help the students gain skills and experience in areas that are relevant to the course. This means that assessors need to be clear what skills the students are expected to develop, and how they can display their competency in those skills. It is also important to consider how the competencies are graded, as on occasion it may be more reasonable for the competency to be represented by a pass or fail mark rather than a grade. Student engagement within a practical assessment can help the student engage with the course itself, as it helps the student connect their

learning to practical applications of these skills. However, this is dependent on the quality of the assessment, as some students may find practical assessments to be unengaging and lack relevance to the educational objectives.

Practical assessments will not be further discussed within this thesis, as they were not relevant to the objectives of this work. However, they are an important aspect of assessment and thus important to discuss when discussing assessment formats.

1.2.5 Selecting an Assessment Format

Students respond differently to different assessment formats, and thus it is important to consider what effects an assessment can have on the students' learning and their approach to a course. One of the most defining characteristics of an assessment is its quality, as regardless of which format is used, if the assessment is poorly constructed the results will not reflect the students' ability.[9,10] The quality of the assessment is influenced by the implementation and the design of the assessment, and is reflected in the assessment's reliability and validity. None of the assessment formats inherently provide a higher quality of assessment.[65-69] This means that regardless of which format is chosen by the assessors, it is imperative that they spend time ensuring that every question asked within the assessment is properly constructed.

When selecting a format, the first consideration of the assessors should be "what is it that we are trying to assess?" as some formats give the students better opportunities to display their ability within a particular area better than others. For example, if the assessors want to assess the research skills of the students, then a written response assessment would give the students the best opportunity to display this ability. Similarly, if the assessors want to test how well students have memorised specific facts within the course, then a MCQ assessment would make the most sense. That is not to say that this is all that these assessments are capable of (as it would be entirely possible to assess research skills using a MCQ and fact memorisation with a written response), but rather how well the assessment format aligns with the educational objectives needs to be considered when selecting a format.[40,60,61] This idea ties in with the concept that specific assessment formats are better at assessing different cognitive levels, which may be defined using any educational taxonomy such as Bloom's or SOLO (Structure of the Observed Learning Outcome).[70-72] It is commonly suggested that MCQs are best at assessing lower cognitive levels (remember and understand), while constructed response are better at assessing higher cognitive levels (apply, analyse, evaluate, and create), where the more detail that the student is required to provide, the higher the cognitive level (e.g. written response is seen to be assessing at a higher cognitive level than short-answer questions).[60,73,74] This idea is true in that it is easier to write questions that assess those cognitive levels using the specified formats, but it does not mean that those formats are restricted to those cognitive levels. The exception to this is the highest two cognitive levels of evaluate and create, as both of these require the students to be able to construct solutions or discuss ideas based on their own knowledge and understanding, which cannot feasibly be done outside of long-answer, written response, or oral formats. However, outside of those two cognitive levels it is possible to assess the other cognitive levels using any assessment format if the items are constructed accordingly.

There are arguments both for and against each assessment format based on their reliability and their construct validity as assessment formats.[13,39,75-78] The reliability of an assessment refers to the chance that the same student ranking would be generated if the students undertook the assessment again (i.e. the repeatability of the results). The construct validity of the assessment refers to how

accurate the measure of performance is within the content being assessed, and thus if the assessment is reflective of the actual ability of the students in the content or if it is influenced by unrelated factors.[79] Often, it is thought that constructed response is a more reliable and valid assessment format, as it requires the students to write out their answers, and thus it is thought to give a better representation of the students' understanding. This is because the validity and reliability of MCQ assessments is thought to suffer due to the potential for students to guess the correct answer. However, because a single MCQ requires a lot less time to answer, more questions can be asked of the students in a timed assessment than could be asked using a constructed response format in the same amount of time, which increases the reliability and the content validity (the degree to which the assessment represents all facets of the content being assessed) of MCQs. This is because as more questions are asked of the students it gives a more accurate depiction of the students' ability across a wider breadth of content. As a result of this, the content validity is naturally higher (as more of the relevant content can be covered in the same amount of time), which can then increase the construct validity of the assessment, as asking more questions of the students decreases the impact of outside influences on the assessment (assuming that the items being asked are well constructed). Theoretically, if the content and construct validity increase so should the reliability of the assessment, as if the assessment is more accurately assessing the content across a wider breadth of content, then the performance of the students should be a more accurate representation of their ability within that content.[4,78,79] In comparison to this, constructed response assessments may have to neglect specific course content due to time constraints that may have been able to be covered in a MCQ assessment. The smaller amount of content covered may hurt or help the results of a student if they are much better or worse in the content area that was not covered by the assessment. While MCQ assessments mitigate this issue by being able to ask more questions within the same time frame, and thus potentially address the issue of missing content within the assessment, they give the students an opportunity to guess the correct answer regardless of their knowledge, which impacts upon the validity of the results.

In theory, this means that both the constructed response and MCQ format can be as reliable, and as valid as each other, assuming that there are no issues within the assessments themselves. The actual evidence for this is conflicting, as some research supports this idea while others refute it.[67,68,80] It is critically important that to compare the reliability and validity of the formats there are no construction flaws within the items, and ideally the items should be reflective of each other (i.e. items with equivalent stems between the constructed response and MCQ items). This is because generally MCQs are more susceptible to flaws in their construction, as they are harder to write; however, constructed response items that contain construction flaws have a much larger influence on student performance.[81] The use of stem-equivalency is to ensure that both assessments are assessing the students on the same content, which ensures that any differences seen are a result of the format rather than student understanding.[67,68,80-82]

It is important to acknowledge that there are advantages and disadvantages to each assessment format, which can make one format more appealing than another depending on the circumstances surrounding the assessment.[39] These advantages and disadvantages, combined with an assessor's own personal experience with an assessment format can cause bias within the assessor when they are selecting an appropriate format for their assessment.[83] In general, an assessor having a preferred assessment format is not an issue, as all assessment formats can provide valid information about the students' ability. However, it is important that assessors can acknowledge when one assessment format is more suitable for an assessment than other formats, regardless of their own bias for or against a specific format. For example, the MCQ format has been shown to be effective in courses

where it has been neglected as an assessment format due to the belief that it would not be able to assess students at the level required for the course.[73] This sort of bias is unhealthy for assessment as it prevents the improvement of the quality of assessments within a course, and thus assessors should be open to any format when they design a new assessment and select the most appropriate format based on the objective of the assessment.

The viability of each assessment format is different depending on the circumstances surrounding the implementation of the assessment, and thus these factors need to be considered when choosing an assessment format. For example, if a course is online, then the restrictions that this places on how assessments can be undertaken need to be considered to ensure that the assessments used are both reasonable for an online course, while still fulfilling the requirements of appropriate assessment. Online courses can have much larger enrolments than regular courses, as the students are not required to live close to the educational institution. These high enrolments would usually limit the amount of assessment done using constructed response formats, due to the amount of time that would be required for the assessors to mark every assessment. In the case of an online course with high enrolment, it is reasonable that MCQs are utilised as one of the predominant assessment formats, assuming that they are used when appropriate and not at the determent of the course. To administer any piece of assessment, the timing and the presentation of the assessment needs to work for both the student and the assessors, and thus the two factors need to be considered when selecting an appropriate format.

1.2.6 Implementing an Assessment

Once the format of the assessment has been decided, the next step is to generate the assessment itself and use it to assess the students. At this stage it is important to have a clear understanding of the desired outcome of the assessment. This refers to both what the assessors will learn from the assessment (e.g. what the students can and cannot do), and what the students are expected to take away from the assessment (e.g. the skills and abilities that are expected of them). It is also important to consider the emphasis placed on each assessment being undertaken, as different weightings of the assessment are a clear indication to the students about what aspects of the course are seen as the most important by the assessors.[9,23,24,84] For example, exams that take place at the end of courses are often weighted quite highly in comparison to other assessments that take place throughout the course. This is a clear indication to the students that the results of the exam are considered by the assessors to be the most important indication of whether the students have met the educational objectives required to pass the course. In contrast to this, assessments that are only given small weightings are viewed by the students as either being assessments that are relatively easy, or they represent an aspect of the course that the assessors deem important enough to assess but not important enough to have a substantial impact on the students final result. Students can use assessment weighting as a way of determining how much time they should be spending on an assessment, and because of this it is important that each assessment weighting is an accurate reflection of what is expected of the students.

Another important consideration when implementing an assessment is when and how that assessment takes place. The timing of the assessment can impact how the students learn the content covered within the course, as it affects the period in which the students will revise the content, and thus it is important that the assessment is aligned with the objectives of the curriculum. Generally, topics are assessed shortly after they have been covered within a course to ensure that the knowledge is fresh in the students' minds, and ideally the assessment will help the students reinforce and retain that knowledge. As an assessment needs to align with the curriculum it

is unreasonable for the students to be focused on one aspect of the course during the semester, with the final assessment taking a different approach. Thus, the idea of curriculum alignment needs to be considered whenever an assessment is planned.

After an assessment has been generated, the next logical step is for the students to undertake the assessment and their performance measured. Depending on the assessment format used, the marking may occur in different ways. In the case of a MCQ assessment it is possible to utilise automated marking to quickly determine the results of the assessment. However, other assessment formats may need to be marked by hand, as the students' answers are subjective in terms of whether they have completely answered the question being asked of them. Depending on the number of students being assessed, more than one assessor may be involved in the marking process, with either different assessors marking different questions or the entire assessment split between the markers. Having more than one marker means that there is the potential for deviations in how each individual student is marked, as the subjectivity of what one marker believes is a complete answer may not match what another marker thinks. As discussed earlier (Section 1.6.3) this is where a tool such as a marking rubric is essential to ensure that all of the students and the markers are aware of what needs to be done in order to show different levels of competency within an assessment. There are ways to monitor if there are significant deviations between assessors marking the assessment, and ways to combat these deviations if they arise. It is possible to moderate the assessors either using an objective third party that re-marks a few assessments from each marker to ensure that the marks assigned are consistent between all the assessors. Alternatively, the assessors may re-mark an assessment that was marked by a different assessor to obtain a comparison between how each of them marks the assessment. In either case, it is important that there is reliability in how each of the assessors mark the assessment in comparison to each other to ensure that none of the students are at an advantage or disadvantage due to who is marking their assessment.

It is also important that the students receive some amount of feedback from the assessment after it has been marked, to provide them with the opportunity to improve if they are unsure of where their performance was lacking.[58] The use of feedback can help to give the students direction, and inform of them of any misunderstandings that they may have had when undertaking the assessment. The timing and the context of feedback is important, as the feedback needs to be able to engage the students in some aspect of their work that they believe is relevant to their performance. If the feedback is received after a significant amount of time since the student completed the assessment, then it is possible that the student has already 'moved-on' from that assessment task. Similarly, if the assessment task was not seen to be particularly relevant by the student (potentially an unweighted assessment or an unengaging task) then the student may be less willing to learn from any feedback they receive from that task. At the bare minimum, the feedback that the students receive will simply be the grade that they received, but this is generally only in cases where the assessment is re-used in some capacity and thus giving out detailed feedback could impact the future use of the assessment. However, the issue with using results as a form of feedback is that it does not give the students any direction on how they can improve upon those results. This could result in the students stagnating, as they do not know where they need to put their time and effort to improve their performance in future assessments.[7,8,85,86] Thus, ideally the students need to receive some indication of what they did well and what they did poorly within an assessment to help them improve. This could come in the form of a completed marking rubric, notes written on the student's assessment that is returned to them, or in some cases it may need to be general feedback given to the entire student cohort about areas that need improvement. It is not necessarily the type of

feedback that is important (although more detailed is better), but that some feedback is given to the students based on their result in an assessment. Thus, it is worthwhile considering how feedback will be given to the students upon the completion of an assessment task.

Another aspect of marking any assessment is reviewing how the students performed compared to what was expected of them. For example, in a regular MCQ or short-answer question there will be some items that the assessors will expect the students to be able to answer easily, as they may assess the students on the basic course content. In contrast, there will be other items that the assessors expected the students to have difficulty with, as they may assess more advanced topics covered within the course. How closely this matches the observed results can be checked after the students have undertaken the assessment by reviewing what items were answered correctly more often by students compared to what items were answered incorrectly more often. This can be used to ensure that the assessment is functioning the way it is intended to, and if the results are not what was expected it can be used to inform what needs to be taught in the future to improve assessment results. There is also the potential that this highlights items that do not perform well due to faults in the items and not due to the students. This might be the result of the item's wording, construction, content, options, or some other factor, but regardless it is important that these questions are identified and either improved or replaced in future assessments. How the quality and the performance of an assessment is reviewed and measured should be considered when the assessment is originally conceived. This is important to ensure that the assessment is fulfilling the task that it was designed for, and is particularly important in assessments that are re-used from year to year to ensure that the quality of the assessment is always high.

The last important aspect to consider when implementing an assessment is how it fits in with the other assessments being utilised in the course. The most obvious concern is that concepts may be assessed multiple times within the course, which may boost the performance of the students who understand those concepts and hinder the performance of students who are not as competent with those concepts. But there are other concerns that need to be considered, like over-assessing the students, which will cause the students to spend more of their time revising for assessments rather than learning the course content.[9,17,23] This means that it is important to select the timing of assessments to maximise the content that can be assessed (e.g. at the end of topics or at the end of the course), and identify the aspects that the students need to be assessed on rather than trying to assess all of the course's content. All of these decisions are made by the assessors, which means that they have a large influence over how the students will approach their learning within the course.[9-11,14,17,19,23-26] This influence extends to all aspects of the assessment (i.e. format, questions, feedback, timing, and implementation), and thus the impact of each decision needs to be seriously considered by the assessors to ensure the best outcomes for the students, and that the best assessments are obtained.

1.3 Factors that Influence Student Outcomes

1.3.1 Student Attitudes toward Assessment

Students' attitudes towards assessments are not based solely on how they perform within the assessment, and as such it is important to consider how the students are going to react to the assessment format in terms of how they revise and retain information.[87] This is because different formats can influence how students adapt their learning approach for an upcoming assessment to improve their performance.[17,19,25-27,88] For example, if the students have an upcoming MCQ

assessment they tend to undertake more surface level revision on a broad range of topics that were covered in the course.[18,27,89,90] This is because the students believe that most commonly MCQs assess general knowledge and comprehension, but as MCQs can assess a broad range of content, the students feel that they need to revise the entirety of the course content. In comparison to this, students who are revising for constructed response assessments spend more time learning the details of a concept or idea, as they expect that they will need a deeper level of knowledge to provide enough detail to answer the questions asked within this assessment format.[18,27,89] These different learning approaches have been shown to result in at least the same amount of retention, but possibly more retention when constructed response style questions are used.[91-95] However, while the assessment format can influence the students' revision, the students tend to already be predisposed to a learning approach.[17,19,25,90,96,97] Despite this, the assessment format and factors such as interest, engagement, long-term goals, and time are considered to have key roles in determining the approach students take to their learning.[98-107] Thus, while assessment format does influence student learning approaches, it should not be seen as the only influencer, and hence assessors should not select an assessment format based on how they believe it will influence student behaviour in regards to their learning approach.

Students tend to have a preferred assessment format, which is usually the format that they feel most confident in and causes them the least amount of stress.[108-110] The preferred format does tend to show some correlation with the ability level of the student, as higher ability students tend to prefer constructed response, while lower ability students tend to prefer MCQs.[111] This is thought to be due to changes in student perception as their ability level changes. Higher ability students prefer constructed response because it enables them to clearly show their knowledge and thought process when answering questions, and they feel that they can achieve better results because of that. Lower ability students prefer MCQs because they know that statistically they always have at least a chance of obtaining the correct answer, whereas this is not true in constructed response formats. However, constructed response assessments generally cause more stress for the students regardless of their ability level.[112] This is because even the higher ability students take comfort in the knowledge that they always have a chance of obtaining the correct answer within a MCQ. In general, students tend to suffer the most stress and anxiety before they undertake a timed assessment. Student stress and anxiety is an unavoidable threat to assessments being an accurate reflection of student ability, as they can influence the student's responses in ways that are unrelated to their ability.[29-31,113] However, once the students begin the assessment, the stress and anxiety tends to leave them as they focus on the assessment task they are undertaking.[32] That being said, the stress and the anxiety suffered before the assessment can cause students to undertake different learning strategies to cope, the most common of which is to study for the assessment and not the course content.[29,30] Students are driven to this approach when they believe that the assessment is not a fair representation of their competency within the course, and thus rather than study the course, they believe they will obtain better results by learning the expectations of the assessment.[10,114] However, the best way to minimise this issue is to utilise assessments that the students believe are fair and valid ways of assessing their knowledge and understanding of the course content.[18] This can be done by ensuring that students understand how they will be assessed within a course, and by building trust with the students that the assessments are designed to measure competency. If the students do not trust either the assessment or the assessor then they may attempt to give the answers they think the assessors want rather than what they believe the answer to be.[14,16,28] Whatever the case, it is important that the students view any assessment to be a fair and valid way of measuring their ability, and that it actually is a fair and valid way of assessing the students.

The students' perceptions of assessment can also be influenced by their attitude toward the course, and, conversely, the assessment tasks can influence the students' attitude toward a course.[20,109,115] Thus, if a student is motivated and engaged with a course then they will likely see assessments as an opportunity to display their learning and identify areas they can improve upon, thus having a positive attitude toward assessment. Conversely, students that are not engaged with a course do not see assessments as a learning opportunity, as it requires them to show some level of engagement with the course, and thus these students are more likely to have a negative attitude toward assessment.[116-119] Raising the students' engagement with the course will likely result in more positive attitudes towards assessments, and thus should improve the functionality of an assessment. The student's perception of their own knowledge and skills, the approachability of the course, and the attendance of the students are all key factors in determining how engaged students are.[116,117,120-124] Increasing all of these factors can help to drive student engagement, but the approaches that will successfully increase these factors varies based on the course and the individual student. In contrast to this, it is possible that students become disengaged with a course after persistent failures and confusion with the course content; however, if addressed, confusion can be a powerful learning tool.[119,125] What this means is that there are numerous other factors outside of the assessment task itself that can influence and affect the students' performance, and not all of these are controllable within the course. This is why poor assessment performance is not always reflective of an issue with the students' ability or the assessment task's validity, and why it is important to consider the variables surrounding the assessment when it is being evaluated.[117,118]

1.3.2 Student Responses in Assessment

Several different factors can influence the way in which students respond to items within an assessment, some of which occur before the assessment and some as a student is undertaking the assessment. Many of the factors that can influence students are discussed above (i.e. the assessment format, revision strategies, attitude toward the course), but when the students are undertaking an assessment there are additional factors that could influence how they respond to the items. The first consideration is students' use of answer strategies when responding to items, as this allows them to maximise their chances of receiving marks that they may have not otherwise received. The strategies employed are dependent on the assessment format and the knowledge of the student. For example, within a MCQ assessment a commonly employed strategy is the use of elimination to decrease the number of options and theoretically increase the odds of success. This means that in a MCQ with five response options, the students may easily be able to identify two or three options that they believe are clearly incorrect, and thus now they theoretically have a 50% chance of obtaining the correct answer as opposed to 20% if all five options were considered. The effectiveness of this strategy is dependent upon the quality of the distractors (incorrect responses) and the ability of the student, as high quality distractors will be hard for the students to eliminate unless they are high ability students.[126] Similarly, in constructed response questions students will often repeat or rephrase the question with small extra details added to gain as many marks as possible with whatever little information that they know or is provided to them. Often if a question is worth multiple marks, this strategy can be used to gain at least partial credit just by beginning to answer the item. These strategies do not refer to students reviewing their answers, nor do they necessarily undermine the quality of the assessment, unless the assessment is poorly constructed and can be taken advantage of by these strategies.[127,128] This is because while students will boost their assessment performance using these strategies, only the highest ability students will have the knowledge required to implement these strategies to achieve the highest level of performance. Thus, any effect that the strategy has on the student's final result is likely to be relative to their ability. If the employment of these strategies influences the students' performance in an assessment

to the point where the reliability and validity of the assessment are in question, then the assessment itself is likely to be poorly constructed and needs to be evaluated and improved upon.

There is a similar but slightly different factor that can influence students' responses within assessments, which is the "test-wiseness" of the student. Test-wiseness is a skill that students build up after undertaking multiple assessments, and it is used by students to obtain the correct answer to questions that they do not know the answer to.[127,129] Students utilise test-wiseness in strategies that are both dependent and independent of the assessment being undertaken. Independent of the assessment format, the students may engage in time-saving, error-avoidance, guessing, or deductive reasoning strategies.[127] Both time-saving and error-avoidance strategies allow the students to maximise the amount of marks that they receive by effectively using their time. This means that they will spend more time on questions that they are confident in and provide short answers to questions that they are less confident in to avoid spending time giving an incorrect answer. Guessing and deductive reasoning are used by the students to give answers to questions that they lack confidence in while still maximising their potential for marks. This is applied differently depending on the assessment format being used (e.g. the elimination strategy within MCQs), as different information is given within the item in different formats. Depending on the assessment itself the students may utilise cueing and intent consideration strategies to answer questions.[127] A cueing strategy is when the students use the way that an item is written and presented as a way of constructing their own answer. This is heavily dependent on the format, as within MCQs cueing is used to attempt to identify the correct option based on the construction of the item and the options. However, in constructed response students look for key points made within the question to inform them how they should approach writing the answer. Intent consideration is when the student does not consider the question being asked but what they believe is required within the answer to the question. This could be based on either knowledge about the assessor(s), or it could be based on previous experience with similar items. Test-wise strategies can be thought of as an extension to regular assessment strategies, as the underlying strategy is the same, but it includes the use of skills that can be improved and are unrelated to the assessment itself. This means that students who are more "test-wise" (either because they have undertaken more assessments or been actively using it for longer) have an unfair advantage over other students. As with other strategies, there is a correlation between the students' knowledge and their ability to implement a test-wise strategy. However, a student that is better able to apply a test-wise strategy (which is an ability unrelated to the course) will perform better than a student with equal ability. There will always be test-wise strategies being utilised within assessments, as the students will use every option available to them to improve their marks. Test-wiseness is not always easily measurable or even noticeable as it impacts students' results in subtle ways due to its dependence on the students' individual ability. This means that there is no reasonable way to account for how "test-wise" individual students are, and thus there is no way to adjust the results to account for an individual's test-wiseness. This means that the best option to avoid test-wise strategies giving students an unfair advantage is to generate assessments that make these strategies ineffective. This means that the assessments need to have a high enough quality that regardless of the strategy employed by the students their result is a reflection of their ability within the course and not influenced by other factors such as test-wiseness.

The students can also be influenced by the assessment item itself, as the item content, familiarity, presentation, demand, and steps can all influence how students respond.[11,26,89,103,130] The content on which the student is being assessed can shape the way in which the students answer the item, as different topics or concepts within a course require different approaches. This has neither a positive

nor a negative impact on the students' responses but affects how the students believe that they need to respond to the item. The familiarity that a student has with an item is due to items that they have been asked in the past, either in tutorials, practice assessments, or in other settings. An item that a student is familiar with is often more easily answered, as they are more practiced in them and understand what is expected of them. However, often familiar items with different requirements can be problematic for students as they often 'autopilot' the questions without much thought for any potential differences. The presentation of an item affects how the students approach the item and their understanding of the task presented to them. For example, if an equation is presented to the students and the item simply asked the students to solve the equation, the students will understand immediately what is required of them. However, if the equation was presented within a large block of text the students would have to extract all the pieces of information and generate the equation before they could solve it. Both items are assessing the students on the same skill (solving the equation), but the way that the item is presented to the students changes how they need to approach the item. The same example can be used to illustrate how 'item demand' affects student responses. In the first example above the item does not demand that much of the students, as it is an equation that needs to be solved, whereas in the second example the item demands much more of the student as they need to generate the equation themselves before it can be solved. The more that an item demands of the students, the more work that they are expected to perform to obtain the answer, which is related to the number of steps that the students need to undertake. All of these factors can result in student confusion, meaning one of two things; either the item is poorly constructed, and therefore the results will not reflective of the students' ability, or the student does not possess the level of understanding required to answer the item. The latter option is reasonable, assuming that the student is confused by more difficult items, but the first option undermines the assessment. Thus, it is important that confusion due to any of these factors does not influence how students respond to items.

The structure of the assessment can influence how it is approached by the students. For example, some assessments are ordered such that the easier items are early in the assessment, and the harder items are towards the end of the assessment. The idea is that it eases the students into the assessment and helps to alleviate some of the stress and anxiety that the students may have at the start of the assessment. As a result of this, the students may take several different approaches to the assessment. Assessors may expect that the students answer every item to the best of their ability. However, the students may spend a lot of time answering the earlier items, as they find these items to be easier and try to guarantee that they obtain full marks on those items, which may result in the students not having the time they require to answer the latter items in the assessment. This in turn affects their overall result, and thus often students that engage with this approach will show a sharp drop-off in obtaining the correct answer later in the assessment. The opposite to this is when the students do not answer the early items to the best of their ability, and instead focusing on the later items as they are more difficult and require more effort. Both of these approaches can occur in any type of timed assessment regardless of how the items are ordered. In assessments ordered by difficulty, both approaches closely resemble a strategy for answering assessments to maximise marks.

The most obvious factor that affects how students respond to items within assessment is their knowledge of the skills and concepts covered within the course. If the students have the required knowledge then, unsurprisingly, they will use that knowledge to obtain the correct answer. However, just because a student believes that they know the correct answer to an item does not mean their response will be correct. When a student incorrectly answers an item that they thought

they knew the answer to, there are some potential reasons for their misplaced confidence. One of those reasons is that the students may not have integrated the required knowledge to answer the item with their existing knowledge, and thus they are not able to apply the overarching concept when answering the item.[131] Students may hold alternative concepts to the actual concept, meaning that even though the student may have complete confidence in their answer, their response may not make sense when the true concepts are applied.[131] This leads to the question of whether the students are uninformed or misinformed. This is an important distinction as uninformed students are simply unaware of what they needed to be doing, whereas misinformed students believe that they are answering the item correctly.[131] To educate an uninformed student they simply need to be taught the relevant concepts or ideas. Educating a misinformed student is more challenging as they need to be convinced that what they are doing is incorrect, and then they need to be introduced to the correct concepts and ideas.[131]

Student responses are driven by a multitude of factors that can be unrelated to the actual ability of the student being assessed.[132] The easiest way to control all of these factors is to use high quality assessments, as high quality assessments can alleviate student confusion and minimise the effectiveness of any answer strategies that the students may try to employ.[77,133] If factors that are unrelated to student ability can be controlled, then assessments will give an accurate measure of student ability, as student performance will be dependent solely on their understanding of the content being assessed rather than any other unrelated factors.[77,134]

1.3.3 Potential Predictors of Success

There are other factors outside of assessment and ability that are cited as having an influence over the students' results.[135] The factors most commonly considered are the students' socioeconomic status, and their previous education experience. None of these factors can be controlled by the assessors, as they are all related to the student's personal lives. The only thing that assessors can do is present the content and the assessment the same way to all the students to ensure that they all have an equal opportunity for success despite any differences that they may have in their personal lives.

The socioeconomic status of a student is used as a way to consider the opportunities that may have been presented to the student, their demographic background and a measure of the level of education that the student's carer(s) have received. A lower socioeconomic status generally corresponds to a lower level of income within the family.[136] It has been observed that students with a lower socioeconomic status tend to perform worse in assessments than students who have a higher socioeconomic status; however, this is not necessarily the fault of the students.[136-138] How the socioeconomic status of the student influences their assessment results could be due to a multitude of factors (availability of help with their study, attitudes towards education, quality of their education, exposure to learning opportunities, etc.), but care needs to be taken when interpreting these sorts of results. The socioeconomic status of a student is a generalisation placed on them based on factors that are out of their control. Thus, while socioeconomic status may be a potential indicator of assessment performance, it is important to remember that assessment should reward students based on their ability, and a student's socioeconomic status is unrelated to student ability.

Students can be divided into several different demographics based on their background. The most common demographics that are used when analysing the results of an assessment are the students' age, gender and race.[83] Much of the research that has been undertaken on demographic differences focuses on gender differences within education. Demographic predictors are not predetermined

based on the gender of the student or when and where they were born. Demographic predictors should be considered as how the environment around the student has shaped their thought process and their approach to learning, which may show some amount of correlation to the gender of the student or when and where they were born.

A student's previous educational experience can influence several factors that may help or hinder them within assessments. The most obvious factor is the knowledge that previous experiences provide the student. However, there are skills and concepts that are taught in some courses that are more useful in future assessments than skills and concepts from a different course. For example, commonly students who study mathematics to a higher level, either at high-school or at university, perform better in science courses than students who did not study mathematics to the same level.[124,139,140] In general, the knowledge and techniques that the students obtain and use in the higher level math course cannot be applied specifically to science courses. This means that the students who studied higher level math must be performing at a higher level due to reasons that are separate to the student's knowledge, and may potentially be the approach and thought processes that they were taught in mathematics. Previous experience also impacts the student's confidence within a course, which in turn can influence a student's motivation for a course.[130,141-144] This could result in students who feel like they performed badly in a previous science course believing that they will not achieve positive results in any science course, or it may not influence the student in any way. In the case that it does influence the students it could be highly detrimental, as the student's self-concept (the student's beliefs about themselves, in this instance the student's belief in their ability within a specific topic/course) shows correlation with their assessment results.[145] This perception could be handled by the student in several different ways: they may seek help and study hard to improve upon their perceived shortcomings, they may give up almost immediately and seek to only pass the course with minimal effort, or the student may avoid taking the course altogether due to their lack of self-confidence. In any of these cases the students' previous learning experiences has a profound effect on the way that they choose to undertake their current learning experiences.

It should be noted that it can be very easy to make generalisations about students based on these sorts of factors. However, while there is the potential to use these factors to identify students that may struggle within a course, they should never be considered to be definitive predictors of success. This is because each factor is unique to every student, and every student is unique, which means that while two students may seem to share the same factors, they may respond to those factors in a different way leading to vastly different outcomes.

1.3.4 Gender Differences

One of the most researched potential influencers on assessments is gender, as it has long been thought that there are significant differences in how male and female students think and behave in assessments.[146-152] The theorised differences can broadly be considered as differences within three categories: course attitudes, skills and abilities, and approaches. One important consideration surrounding the theorised differences is whether or not these differences (assuming differences truly exist) are caused by genetic differences between males and females resulting in differences in maturity and development, or if the differences are the result of different experiences that males and females may have. This question of nature versus nurture is not easily answered, as determining the root cause of any variation observed in the performance of male and female students is difficult. That being said, it is unlikely that any identified gender bias will ever be solely attributed to either nature or nurture, as both of them influence how an individual performs.[153,154] The reason that this is important, particularly in education, is because any differences due to nurture can be addressed to

some extent, whereas the inherent nature of the students cannot be changed. What this means is that if the cause of gender differences is due to student experiences, then one way to overcome that is simply to give all the students a similar experience that will help them learn a particular concept. For example, one study within physics identified that there were significant differences in how well male and female students performed in questions about projectile motion.[155] It was suggested that this difference could potentially be due to the relative amount of real-world experience that male and female students have with projectile motion (e.g. throwing and catching objects).[155] This means that the difference in the results of male and female students was attributed to the effect of practical experience influencing the students' ability. In a follow-up study the researchers focused on ways that they could decrease the gender gap through the use of teaching interventions.[156] The researchers found that getting the students to undertake a physical example of projectile motion and then work through the physics of what was occurring both visually and mathematically increased both male and female students' results on an item that had previously shown gender bias. Most importantly, despite it increasing the performance of both genders it decreased the gender gap observed within the item significantly. What this highlights is that there are ways of improving the performance of the students to decrease any observed gender gap, but what was important was that the intervention was able to link multiple ways of understanding a concept. In this example the visual, mathematical, and verbal representations of projectile motion were all linked together within the same activity to ensure that students could explain the concepts in whatever method of understanding worked best for them.

When considering the attitude that male and female students may adopt towards a course, it is very important to consider what the students expect of themselves and what others expect of them. Very commonly students are told, and believe, that male students are better at courses like physical science and mathematics, whereas female students are better at courses like English and natural science.[149,153] This can then affect the courses that the students believe that they can succeed in, and in some cases can cause the students to undergo self-selection away from a course.[157] Self-selection is when students select courses based on what they believe they can succeed in, will enjoy undertaking, and further their goals; which means students engaging in self-selection are unlikely to choose courses that they expect to perform poorly in.[33] The idea that female or male students are more likely to self-select into or away from specific courses is seen more clearly at higher levels of education. At the higher education level, there is a clearer disparity in the number of male and female students in particular courses, and at least some of this difference in enrolment can be attributed to self-selection due to societal stigma.[158-163] That means that either male or female students (depending on the course) may not enrol in a course they otherwise would have because it is seen as either a very male- or female-dominant course. An example of this is nursing students and physics students; historically, nursing has been perceived as a feminine profession and physics a more masculine profession. That is not to say that there are no male or female students participating within these courses, but the cohort of students tends to be skewed in favour of one gender. This view of the course can also influence the results of the students within the course. For example, female students are perceived to perform worse than male students within physics, and if female students believe this then it can result in them performing worse than they otherwise would due to how this stigma influences their self-confidence.[164,165] This means that how a student views a particular course can influence them in two ways: firstly, it may impact their decision as to whether they enrol into a course, and secondly, it can affect their academic results within the course based on their own self-concept of how they should be performing.[142,145,166-168]

Another much-discussed topic surrounding the issue of perceived gender differences is whether male and female students have better affinities with different skills and abilities. This is concerned with whether one gender is actually just better (or worse) at some courses than the other gender. For example, the majority of research suggests that female students are better at verbal tasks, whereas male students are better at visual-spatial tasks and applying knowledge in real-world contexts.[148,149,154,169,170] However, as these differences are not seen in every assessment, it does raise the question of what is making these differences appear. The most obvious theory is that male and female students are simply better at those respective tasks than the other gender. Another theory could be that male and female students are better at the content that is often asked in those style of questions. However, if one of those theories is true, then why are gender differences only seen in some instances and not always? This is an example of why any potential gender differences are a complex issue and cannot be solely attributed to one factor, and hence it is extremely difficult to predict which items will display gender differences before they are used.

The last broad consideration of gender differences is how the students approach their learning and assessments, as it has been suggested that self-discipline and assessment format can play a role in gender differences. The idea that self-discipline impacts assessment results is reasonable, as students who put more time and effort into a course should be rewarded for that effort in assessments. One way of explaining why female students tend to achieve higher grades than male students, but then perform as well or worse in standardised tests, is that female students tend to show more self-discipline throughout the semester, which can help students achieve higher results in regular assessments but is not as helpful in standardised tests.[171] A self-disciplined student is one that will put off a more pleasurable activity in favour of putting effort into a specified goal or task, in this case undertaking study or completing assignments. By prioritising these activities, it is reasonable that a self-disciplined student will perform better on these tasks. However, self-discipline will not necessarily provide any benefit within a standardised test, outside of extra study that the student may have undertaken throughout the semester, as the very concept of standardised tests is that it removes all of those factors and assesses the student purely based on their performance within the timeframe given for the assessment.[171] In contrast to this is the idea that female students perform worse than expected on standardised tests because the format favours male students.[167,168] Without discussing some of the issues of standardised testing, in general, the reason that standardised tests are thought to potentially favour male students is because they tend to have a significant number of MCQs. MCQs tend to be thought of as simple questions that usually require the students to use their knowledge to answer a straightforward question related to the course content. Occasionally the question may require students to interpret diagrams, equations, or figures that are presented within the question. There is generally less, or the same amount of comprehension involved in a MCQ than constructed response, as the items are usually short, nor do either of these formats test the students' verbal abilities. Based on what is thought to be the strengths of male and female students the MCQ format appears to favour the strengths of male students and not the strengths of female students.[172-177] However, the generalisation of female students being more self-disciplined and the strengths of male and female students are not true for all students. This makes trying to balance how much of a result is influenced by outside factors and how much is purely a result of student ability difficult, if not impossible in some instances. Therefore, the best approach to understanding gender differences is through the investigation of assessment performance, where the goal is to minimise or remove any influences that are unrelated to student ability.

Gender differences are an extremely complex issue within courses and assessment due to the number of factors that need to be considered when trying to uncover the driving force behind any differences. The most important aspect of gender differences is having a plan to determine whether they are present within the assessment. It is not enough to simply look at assessment results and conclude that there are or are not gender differences based on the average results of male and female students. This is because factors like self-selection and course bias could be influencing the results of the students such that it should not be expected that the two groups will have an identical distribution of results.[173,178,179] There needs to be deeper analysis of the assessment employed to ensure that any items identified as containing gender bias truly show a statistical difference between how male and female students performed which is not due to variance within the population. Minimising or, ideally, removing differences in performance between genders is an important educational objective to ensure that there is no unfair advantage within any course or assessment, a goal that extends beyond just the subgroup of gender. It will also help to lessen the stereotypes surrounding some courses and careers that are perceived to be male- or female-dominated areas, as it could help students with their confidence and ensure that student perceptions of a course is not a driving factor in student self-selection. Since differences in gender performance first became an issue that was discussed in education, improvement has been made in ensuring that the differences in performance and perception are minimised.[151,158-161] This means that with continued work and effort it should be possible in the future to reach a point where gender differences are no longer a concern within education.

1.4 Multiple-Choice Questions

1.4.1 The Multiple-Choice Question Format

Multiple-choice questions (MCQs) came at a time when enrolment in education was increasing and there was a need for an assessment format that could be used in large scales and be marked quickly. The MCQ format was first used in 1914 by Frederick Kelly, who at the time saw the need for a piece of assessment that was standardised with predetermined answers to avoid bias in the marking of assessments.[180] It was crucial to Kelly that there were no ambiguities within any of the questions being asked and that they had either completely right or completely wrong options to avoid the potential for confusion. This took the power of judgement away from the assessors and enabled standardisation, which allowed for assessments to be graded faster and on a larger scale. The invention of automated marking for MCQs (launched commercially in 1937) made using the MCQ assessment format even more appealing to assessors.[181] The importance of MCQs in assessing large groups of students quickly is still extremely relevant to this day. However, using MCQs as the sole assessment format is not advisable, as it cannot be used to probe student thinking.[34,42,49] While assessors still disagree over the use of MCQs, some believe the format to be flawed and others believe it to be highly effective, it is generally agreed that a mix of assessment formats is the best way to assess the students.[34,39,41,42,49,60,73,76,151,182]

One of the most consistently raised issues with the MCQ format is that they are only capable of measuring the lower levels of Bloom's taxonomy (or any education taxonomy).[49,70] While it is true MCQs often struggle to assess the higher levels of these taxonomies, a well-constructed MCQ is able to assess all the levels of the taxonomy that can be assessed using short-answer questions.[65-67] Thus, when constructing a MCQ it is important to reflect upon what information is being presented to the students and what is required from them based on that information, as it is the item's construction that will determine the level that is required from the students to answer the item.

1.4.2 Construction Factors within Multiple-Choice Questions

A MCQ is constructed around four basic components: the stem, auxiliary information, answer option (key), and distractors. Each of these components contains factors that can influence the responses of the students. This means that when MCQs are constructed, each of these components needs to be considered both independently and together to ensure that the potential for any issues within an item is minimised. It is impossible to know for certain how a student will interpret an item; however, by reviewing the construction of an item before and after it is used within an assessment it is possible to obtain an understanding about how each of these components can influence students' approaches to items.

The stem of a MCQ needs to pose a question to the students that, ideally, they can answer using only the information provided within the stem and any auxiliary information provided. The stem of the item needs to set the intent, content, and the context of the rest of the item. For example, if the intent of the item is to test students on a particular concept, then it should be clear within the stem that the concept is integral to the item. If the students need to look at the answer options to obtain the context or the content of the item, then the stem has not included all the required information. The wording and the presentation of the stem can influence the student's interpretation of the item, which can then impact the student's thought process and the method they use to obtain the answer. This can have both a positive and negative impact on the assessment, as on occasion it might be useful to present the stem in such a way that lower ability students are likely to misinterpret its intention. However, it is important to remember that the students are being assessed on the course content, not their ability to read and understand stems that were deliberately constructed poorly to confuse the students.

Any auxiliary information provided to the students needs to be relevant to the item and should not be used to add unnecessary steps to the items. This is because auxiliary information can complicate items, and depending on the content being assessed it may be more applicable to use a different assessment format. For example, an item that provides a graph as auxiliary information and asks the students to identify the trend within the data is reasonable for a MCQ format. However, if instead the item gave the students the raw data and required them to plot the graph themselves before interpreting the trends then this would be better assessed using one of the constructed response formats. The reason for this is two-fold: firstly, it provides the opportunity to assign partial credit to the item based on the working that the student shows in their answer. Secondly, a short-answer format will provide more information about how the students used the auxiliary information to answer the question. While the distractor options within a MCQ might represent different errors that the students are expected to make, there is no guarantee that the students selected that particular option because of the error it is associated with. It is also important that the auxiliary information is referred to in some way within the assessment, either within the stem of the item that requires it or at the beginning of the entire assessment task if the information is something that needs to be used when answering multiple items. This is because the students need to be aware of all of the information provided to them to answer the items to ensure that the assessment is fair for all the students, and produces results that are solely dependent on the student's ability.

The answer option (also known as the answer key) within a MCQ need to represent a non-subjective correct response to the question posed within the stem. The answer option should always be embedded in facts, and thus any items that rely on the opinions of either the students or the assessors need to be placed in a different assessment format to give the students an opportunity to

explain their reasoning. In addition to this, it is important that the answer option does not stand out from any of the distractors used. There are a variety of strategies that students employ in an attempt to 'game' credit within the MCQ format, one of which is to identify the answer option based on how it is different to the other options. The most common factors that students use are option length and similarity to the stem, as typically the longest option that contains similar wording to the stem is the correct option in a poorly constructed item.[44,183,184] This can easily be avoided by changing either the answer option or the distractors. However, it is important to remember that students will use strategies outside of their knowledge to obtain the correct answer, and thus the presentation of the answer option needs to be considered to avoid any cueing effects. Another concern of assessors is the positioning of the answer option within MCQs, but the best way to counteract this concern is either to use logical ordering (e.g. if the answers are numerical place them from smallest to largest) or to simply randomise the ordering of the options. The research analysing the positioning of the answer option has largely found that while students will attempt to rationalise the answer positioning of the options within a MCQ assessment they are unlikely to change their answer as a result of it.[185-187]

Distractors must represent plausible alternatives to the correct answer to ensure that students are able to display their knowledge and understanding of a topic. It is not always possible to have plausible alternatives if the items assess basic concepts that are expected to be well understood by most of the students. The two best methods to generate plausible distractors is either using expert knowledge to generate possible alternative concepts, or by using students' own incorrect responses to similar questions.[44] The first method relies on assessors creating options based on their knowledge of the topic and the students that they believe represent student misconceptions. The second method requires asking the students several short-answer questions, either in a summative or formative setting, and generating MCQs based on the short-answer questions and using the incorrect student responses to create distractors. It is also important to consider the actual number of distractors used within each item. A higher number of distractors mathematically decreases the chances of students guessing the correct answer. For example, having five options instead of four changes the odds of guessing the correct answer from one in four (25%) to one in five (20%). As a result, usually an item will include three or four distractors (four or five option items), but it is not uncommon for one or two of these distractors to be highly dysfunctional (seeing almost no selection outside of seemingly random guesses). This is because the more distractors that are used, the harder it is to make them all plausible alternatives to the correct options, and thus it is not uncommon that students will be able to identify these weaker distractors in an item and ignore them. If generating enough plausible distractors to create a four or five option item is a concern, it is worthwhile considering decreasing the number of options presented within each item, as it has been shown that three and even two option MCQs are just as valid a way to assess the students (and potentially even more so), providing the distractors are of a high quality.[126,188-190] While this may at first seem counterintuitive, if the dysfunctional distractors are considered irrelevant, as most students simply ignore them, then simply removing them from the item has no real influence on how the students approach the item. By decreasing the number of distractors within every item there are two potential net positives: firstly, this helps to ensure that every distractor is a plausible alternative and not simply making up numbers. Secondly, by decreasing the number of options, the amount of time it takes students to answer an item decreases, which theoretically means that the students could be expected to answer more items within the same time frame and thus validity and reliability of the assessment can be improved through the addition of high quality items.[189,191,192] Distractors should not be considered as options that are presented to the students to trick them into selecting an

incorrect answer; but rather they should be viewed as a method of uncovering common misconceptions and errors made by the students that can be used to evaluate their performance.

1.4.3 Types of Multiple-Choice Questions

The MCQ format that is most commonly used is the single-best response format, where a question is posed to the students within the stem and they must choose the answer option that best answers the question.[34,193] However, there are a number of different formats and modifications to formats that are classified as MCQs.[39,194-198] The other formats are: true-false, multiple-true-false, and matching options. It is also possible to use MCQs in adaptive (alternative) choice assessments or to have students assign their level of confidence to each option presented.[53,131,199] Each of these formats have their own advantages and disadvantages, but so long as the items are well constructed and the assessment is implemented well, each format can function as a valid and reliable assessment.

The single-best response format is the most commonly used because assessors tend to have had the most experience with this format, and hence feel the most comfortable using that format. The reason that this format is used so prolifically within assessments is because it is a highly flexible format that can be used to assess multiple levels of student understanding, over a broad range of topics, in a quick and efficient manner. The analysis of the results can also be insightful for the assessors to gain an understanding of what the students struggle with, as well as commonly held misconceptions within the student cohort. These advantages are highly lucrative for assessors who need to assess large cohorts of students. The main reasons not to use this format are that it can be difficult to write the items, particularly items that assess higher order thinking, and it cannot assess the students on their ability to present their own understanding. The first disadvantage is generally handled using an item bank, where the items are written once and then they are re-used over multiple assessments until they become irrelevant to the course. The second disadvantage must be overcome using other assessment formats, as regardless of the MCQ format used the students will never have to write their own answer to an item. Many of these advantages, and all the disadvantages, are true of all the multiple-choice formats.

The true-false format was used quite regularly within assessments in the past, but it has been used less more recently due to concerns over the percentage chance for students to correctly guess the answer and due to the difficulty of writing quality items for the format. Theoretically, the students have a 50% chance of obtaining the correct answer to any true-false question, as there are only two options for them to choose between. This causes some concern about the validity and reliability of the format, as theoretically if students obtain the correct answer in 40% of the items through their own ability they can guess the other 60% and assuming they guess correctly 50% of the time their final grade will be 70%. Therefore, to counteract guessing there needs to be many items used in a true-false assessment to distinguish between high and low ability students. Another noteworthy aspect of true-false items is that they tend to only measure low cognitive levels, meaning they are best used when assessing a student's knowledge and comprehension abilities. True-false items can sample broadly due to the low response time required, and they can effectively diagnose basic student misconceptions. This means that usually true-false items are more effectively used within formative assessment tasks to engage the students in the content or to determine the students' base knowledge in a topic.

There is a MCQ format that combines the singe-best response format with the true-false format called the multiple-true-false format. The items used in this format are constructed almost the same

as they would be in the single-best response format except the students need to indicate whether each option presented to them is true or false. The reason that some assessors prefer this format over the single-best response format is that in the single-best response format, the assessors only gain information about the students' preferred answer and receives no information about what the student thought of the other options.[200] For example it is possible that a student may select the correct answer but still believe an incorrect option to be true, or conversely they may select an incorrect option but still believe the correct option to be true. The multiple-true-false format can be used to determine if the students have a complete understanding of the concepts within a question, or if they hold partial misconceptions. The use of this format also does partially remove the potential for student answer strategies, as these strategies usually rely on the students comparing the options presented to them. However, in the multiple-true-false format the students need to evaluate each option almost completely independently from the other options (although it is likely that at least a few questions will have contradictory options). Because this format combines the single-best response and true-false formats, it does retain some of the disadvantages of both formats. As all the options need to be answered true or false, it limits the types of questions that can be asked and the level of knowledge that can be assessed. Compared to the true-false format, guessing is not as significant an issue, as guessing can be easily observed when analysing the student's responses, but it is not as time efficient either. There is also a higher level of analysis required by the assessors to understand the students' answer patterns and what that implies about the students' understanding. However, the potential to gain information about the students' mixed and partial understanding of topics is a significant advantage this format has over other MCQ formats.

The last multiple-choice format is matching questions, which requires the students to match an answer, phrase, symbol, term, function, effect, operation, or principle to their correct counterpart. This often is used as a fill-in-the-blank style question where something has been omitted from a sentence and the students need to complete the sentence using one of the options presented to them within a list of possibilities. This format is an effective method for assessing students on their recognition of relationships and associations, but they are not well suited to higher order thinking as they cannot assess any type of interpretation, judgement or application of knowledge.[45] While these items tend to be easy to construct and score, they are extremely vulnerable to student answer strategies. This is because matching can be performed through memorisation and association rather than understanding.[45] Additionally, if the number of options presented is the same as the number of blanks then students will simply utilise the options that they know to be correct and then fill in the blanks using whatever options are left. This means that these style of questions are generally much more effective in lower education levels where it is important that the students are able to recognise relationships and associations.[45]

MCQs can be used to develop adaptive or alternative choice assessment tasks, whereby the items that are presented to the students change depending on their success in previous items. This can be used to derive a more accurate measure of the student's ability, as the students will continually be asked items of varying difficulty until the limitations of their ability can be found. The reason that MCQ formats allow for this type of assessment is due to their objective nature, meaning a response is always either correct or incorrect (assuming that the MCQ was constructed correctly). This means that the assessment can be marked as the students complete it, and thus the items that are presented to them can be based upon their previous results within the assessment. For adaptive assessments to be undertaken in an efficient manner they need to be undertaken on a computer to allow for the answers to be marked immediately, resulting in a seamless transition from one item to the next. This means that adaptive assessments usually take much longer to prepare than a typical

MCQ assessment. This is because more items than usual are required, as some items will only be seen by the highest and lowest ability students, and the entire assessment pathway has to be mapped out so that the next item presented is based on the previous results of the student in the assessment task. However, after the initial setup is completed the rest of the assessment should run in an automated manner, as the assessment will be marked as the students complete each item. Any feedback given to the students can be implemented within the assessment beforehand and received upon the completion of the assessment. This means that while there is a lot of setup required before the assessment takes place there is very little work that needs to be done by the assessors once the setup is complete.

Another approach that has been taken in MCQ assessments is for students to assign their level of confidence to each option that they select. This is most applicable to the single-best response format; however, it can be applied to every format, but what value it adds to the format should be considered. The assignment of confidence can be used either as a factor that is considered when marking the assessment, or it can be used as a diagnostic tool to determine how confident the students are in their knowledge.[131] In the case where assigning confidence is influential in the students' marks, it means that rather than the students responding with only a single option, they may instead give their level of confidence in any number of options. If the student is confident that they know the correct option they may assign that option 100% (i.e. they are 100% confident that that option is correct), but if the student is unsure about two or more options they may decide to split their confidence either evenly between those options or perhaps they might favour one option that they believe is more likely to be correct. If confidence is merely used for student diagnostics, then it can inform the assessors about the predominance of student misconceptions within the course. The students will always select the option that they are most confident in, but if they were 60% confident in the answer they selected and 40% confident in another option this could inform the assessors about the students' thought process. The obvious advantage to students assigning their confidence to each option is that it provides lots of extra information about the students' thought process when they are answering the question.[53,131] However, there are two major issues with students assigning their confidence; the first is that confidence is a somewhat subjective measure from the students (i.e. what is the difference between a student that is 60% confident and a student that is 70% confident?), and secondly, marking based on confidence can be complicated and time consuming.[53,131] The issue with confidence-based marking is that depending on the scoring method used, students can 'game' marks from the assessment format by strategically assigning confidence based on what will give them the most marks rather than their own understanding. This is because unlike the single-best response format where the students have to select one option within this format, the students can eliminate the options until only the ones that they believe are plausible remain and then assign equal confidence to those options. This means that students will always be able to receive some credit for their response so long as they did not eliminate the correct option from their plausible options.[53] There is also the potential that the students who are willing to back their highest confidence will be awarded more marks than students who express their uncertainty.[53] For example if a student is 70% sure that one option is correct they may either express that in their answer, or they may decide to show complete confidence in their knowledge and say they are 100% confident in that option. Theoretically the student believes they have a 7 in 10 chance of being correct, and if that was true then mathematically expressing that they are 70% confident is a better strategy for obtaining consistent marks. However, a student being 70% confident in an answer does not mean that it is actually the correct answer 7 out of 10 times, as the student might be working under a misconception or be influenced by their own self-confidence in their abilities. Therefore, some students may prefer to assign their confidence, while others may wish to place

their full confidence in whatever answer that they believe to be the most plausible.[53,131] The problem is not that one strategy is more efficient or awards more marks than the other; the problem is that this has introduced a factor completely unrelated to the student's ability that can affect the results of the students. As a purely diagnostic tool, this is unlikely to be an issue, as all of the students will have to choose only one option as their answer regardless of how confident they are in the other options.[53,131]

1.4.4 Scoring a MCQ Assessment

All the MCQ assessment formats described above have been successfully used within courses to measure the ability levels of the students and provide feedback to the students about their strengths and weaknesses. However, the way that an assessment task is scored can have a large impact on how the students approach the assessment and what information the assessors obtain from that assessment.[201] For most MCQs the scoring is straightforward: one mark is awarded for a correct answer and no marks are awarded for an incorrect answer. However, some MCQ assessments implement negative marking, "score correction" (i.e. to account for guessing), partial marks (based on the response chosen), or a cut-off point for passing. Negative marking is also relatively straightforward in that if the students select an incorrect option they lose a mark from their overall score. This has a flow-on effect to the rest of the assessment where students are advised to either leave questions answerless if they are not confident in their answer, or select the option of "don't know" which gives a mark of zero for that question. However, there are some concerns that negative marking may reward risk-taking students and incorporate a systematic bias against risk-adverse students, as often an educated guess it better than providing no answer.[202] The most basic score correction uses the number of correct answers given (R) minus the number of incorrect answers given (W) divided by the number of incorrect options in each item (C [number of options] - 1) to give the approximate number of correct responses due to guessing, as represented in Equation 1, but there are other score corrections that can be used.[203,204]

$$Corrected\ Score = R - \frac{W}{(C-1)} \qquad \textbf{Eqn. 1}$$

Equation 1 works under the assumption that the chance the students obtained the correct answer through guessing is solely based on the number of options within the item.[203] For example, if there were five options in the item this formula assumes that for every five items the students guess they obtain one correct answer. Thus, under this assumption every four incorrect answers given by the student means that one mark is taken away from the student's score to account for a correct answer obtained through guessing. Another way that the score obtained on a particular item could be calculated is to award partial marks based on the response chosen by the students.[54] Within each item there would need to be one correct answer, a few options that indicate partial understanding, and some options that are completely incorrect. This allows for students with a partial understanding of a topic to gain marks for that understanding to help assessors differentiate between high, medium, and low ability students. A cut-off point for passing is used so that the students are required to have at least some level of knowledge to show competency within the assessment task, as simply guessing the answers will never award the students enough marks unless they are lucky (which then calls into question the validity of the assessment if students are able to show adequate performance in this way). Each of these scoring methods complicates the assessment format but depending on the assessment the assessors may prefer to complicate the assessment if they feel like it will make it more valid and reliable overall. Regardless of the scoring

method used within the assessment, it needs to be clearly described to the students beforehand to ensure that they understand what is required of them within each assessment task.

1.4.5 The Use of MCQs as an Assessment Format

When using any assessment format it is important that the assessors are confident that the results are reflective of what they are trying to assess, and that similar results would be obtained by the students if they were assessed on the same content again. This is represented by the assessment's construct validity and reliability, respectively. The construct validity of an assessment is used to determine if the performance of the students is an accurate reflection of their ability in the content being assessed.[79] If an assessment is reliable it means that if the students resat the same assessment, with no knowledge of their previous attempt, they will obtain the same result consistently. These two concepts are critical to any assessment format, not just MCQs, and as such it is important that the assessors are able to confidently state that all of the assessments used within a course are a valid and reliable way of measuring student performance. However, while the format of the assessment can have some influence over the validity and reliability of the assessment, it is the items used within the assessments that determine if an assessment is valid and reliable. This means that any assessment format can be valid and reliable (assuming the format fulfils the assessment requirements) if the items used are of a high quality. This is due to how the validity of an assessment is evaluated, and how reliability is calculated. The best way to identify issues with an assessment's construct validity is to evaluate how much construct-irrelevant variance is present within the assessment. This represents any variables that are not controlled within the assessment that can affect the performance of the students but are completely unrelated to the content being assessed. These variables are often introduced within the items used in an assessment, as outside of the potential for students to guess the answers, the format itself has no way of introducing these variables. It is possible that these variances are introduced through a lack of curriculum alignment within the assessment, which is represented by the content validity of an assessment. Content validity is used to ensure that an assessment is assessing all facets of the course content that is expected to be present within that assessment, and that no one area is over-represented based on the importance assigned to it within the curriculum.[79,205] This is best evaluated by reviewing assessments both before and after they are used to ensure that everything within the curriculum is assessed to some extent. Another way of considering the validity of a piece of assessment is how well its results align with the results obtained in other assessments undertaken within the course. This is evaluated by the criterion validity, and is best calculated through correlation measures between the assessment being analysed and any other performance evaluation that the assessment is expected to show similar results to.[79] This may involve the comparison between results obtained in one aspect of the course (e.g. a student's performance within one practical compared to their overall practical performance, or comparing a student's performance in multiple tutorials undertaken throughout the semester), or comparing the results of an assessment to the final result obtained within the course. Assessment reliability can be determined by the correlation between two alternate forms of the same assessment, the correlation between the results of the same assessment taken at two separate times or based off the consistency of the results within a single assessment. Comparing two assessments has its limitations, as the differences between the two assessments and the period between undertakings of the assessments are factors not considered within the correlation calculation. Instead, reliability is generally measured as the internal consistency of the assessment by comparing the performance of the students across all of the items within the assessment to determine if there is consistency in the types of items that students are answering correctly and incorrectly. If an assessor is concerned about the validity and reliability of their assessment, rather than immediately changing the assessment format, they should first analyse

the items that they utilise within their assessments to ensure that they are performing the way that they are expected to.[65,73,81]

An important aspect of using any assessment format is knowing the limitations of that format and where it is important to assess the students in different ways to ensure an accurate measure of ability is obtained. It is well-known that MCQs are less able to assess the highest cognitive levels of evaluate and create, which require the students to create new ideas or solutions and be able to explain and summarise judgements based on ideas and concepts taught within the course.[34,39,42,43,49,53,66,75,82] These levels can never be assessed within MCQs as they require the students to be able to demonstrate their own thinking, something that can arguably only be done using long-answer or verbal assessments.[70-72] However, if the items are constructed well it is possible to assess all the cognitive levels lower than those two (remember, understand, apply, and analyse), as the students do not have to directly show their thinking to display these levels.[206-208] That does not necessarily mean that all of those levels should be assessed using only MCQs, it simply means that MCQs are capable of assessing all of those levels. For example, generally MCQs are used to assess the students very broadly across the course content, and then constructed response items are used to assess the most important concepts and ideas within the course. Not only will this make all of the course content relevant to assessment, but it will also give students the opportunity to clearly display their understanding of the most important concepts, which in turn provides the assessors with more information about the students' ability. The assessment format should be selected based on the requirements of the course, which means that in some cases MCQs can be highly effective at measuring a range of cognitive levels, and in others they may simply assess basic recall of knowledge. However, it is important to remember that the cognitive levels assessed within the MCQ format is dependent upon the items used within the assessment and not solely the MCQ format.

One highly undesirable trait of any assessment is the idea of construct-irrelevant variance, which is when the student's answers to the assessment are influenced by factors that are completely unrelated to the student's ability.[134] The reasons that this may occur varies between assessment formats, but typically the most common reasons are due to poor construction of the items, answer strategies, answer recognition, and test-wiseness. All these factors are a much greater concern within the MCQ format, as the students are selecting between different options rather than starting their answer from nothing. This means that any effects that are unrelated to student ability and influence the answer of the student will have a larger impact on the student's results within an MCQ assessment than they would in any other assessment format. This is a significant concern for assessors, as they want the results to be purely reflective of the student's ability, and thus potentially some assessors will disregard MCQs as an assessment format due to concerns with construct-irrelevant variance. However, it should be noted that many of the issues with construct-irrelevant variance can be related back the items being asked to the students. For example, there is concern that if a MCQ only requires simple recall of basic knowledge then it is possible that the students may be able to obtain the correct answer by recognising it within the list of options presented within the item rather than answering the item based off their knowledge of the content. If the students truly do only obtain the correct answer due to recognising the answer option, then obviously this does present an issue with the nature of the MCQ format. However, there are several potential changes that can be made to improve such an issue within a question. The recognition of the answer could be prevented by either changing the way that the question is written such that the same knowledge is still required of the student, but the recognisable answer is removed. There is also the potential that the problem is not the option itself but the distractor options, as they do not

provide plausible alternatives to the correct option. If the question cannot be re-written and there are no plausible distractors, then either the question is not suitable for the MCQ format, or potentially the actual content of the question is likely to be fairly easy for the students, meaning the format is not the issue. The same logic applies to issues with test-wiseness and answer strategies, where often it is easy to minimise or negate their effectiveness simply by addressing those concerns within the construction of the question. This means that while there is some construct-irrelevant variance that is unique to the MCQ format, construct-irrelevant variance does not occur within every MCQ item and it is possible to minimise or remove its effect by careful construction of the items used.

It is important with any assessment format that the assessors know what they expect from the assessment, both in terms of how they expect the students to perform, as well as the information that they expect to obtain about the students from an assessment task. The MCQ format offers assessors a time-efficient way to measure the ability of the students across a broad range of topics, but if utilised incorrectly it does have several drawbacks. The most severe drawback to MCQs is their susceptibility to construct-irrelevant variance, either because of the student taking advantage of the assessment or due to issues with the construction of the question. Despite this, as long as the drawbacks are taken into consideration in some way (either by using other additional assessment formats, constructing quality questions, or adjusting the scoring method), MCQ assessments can be used as an effective method of assessment in almost any course.[73,209,210]

1.4.6 Taxonomy of a MCQ Assessment

There have been a number of different taxonomies generated in order to classify the level of understanding required in order to answer assessment items.[71,211-215] These taxonomies are quite broad, which is both an advantage and a disadvantage. The broad nature of these taxonomies allows them to be applied to any level of education, within any topic, and to any assessment format. But this means that they do not account for factors that are unique to each topic and assessment format.

The most commonly used taxonomy within education is Bloom's revised taxonomy,[70] which bases educational outcomes on three different domains: the cognitive domain, the affective domain, and the psychomotor domain. The cognitive domain is based on the development of students' intellectual abilities and skills. The affective domain relates to the student's interests and attitudes, and the developmental adjustments to their thought process. The psychomotor domain is related to the student's physical skills, and is the often the least relevant domain as it plays almost no role in some educational objectives. To classify items, only the cognitive domain needs to be considered, as the other domains relate specifically to the students themselves. This can be used to describe the intended behaviour of the students for them to correctly answer an item. The intended behaviour of the student does not always represent the actual behaviour of the student, as some students use different skills or approaches to answer items. The deviation of the student from the intended behaviour is problematic in ensuring that the students have developed the intended skill. However, these are the sorts of considerations that need to be made when constructing and evaluating items to ensure that the students are being adequately assessed in all the relevant areas.

The cognitive domain within Bloom's revised taxonomy is based around six different groups that require an increasing level of intellectual skill and understanding, which is usually described as different orders of thinking. These groups are: remember, understand, apply, analyse, evaluate, and create. Each of these groups then contains its own subgroups that further describe exactly what

level of skills, understanding and order of thinking is required to answer items within its group. These groupings have been used to classify the level of understanding that is required to answer an item at all levels of education. They have also been used to argue for and against different assessment formats. Commonly, it is claimed that MCQs are only effective at assessing the first three levels of the taxonomy (remember, understand, and apply), and to reach higher levels the short-answer or long-answer assessment formats are required. Rather than restricting the level of assessment to specific formats, this taxonomy should instead be used to evaluate each item on its own merits, as this will allow for a more accurate classification of the items being asked of the students.

Another common taxonomy used in education is the Structure of Observed Learning Outcomes (SOLO), which is used to describe a student's level of understanding of a subject based on increasing levels of complexity.[72] These levels of complexity are: prestructural, unistructural, multistructural, relational, and extended abstract. As the students increase their level of understanding they move through the different stages, starting with knowing nothing they learn the basics of a subject, then they learn different aspects of the within that subject. Once they understand different aspects of a subject the next step is being able to link those aspects together in different ways to explain and analyse outcomes. Finally, once a high level of understanding has been obtained it may be possible to apply their understanding of that subject to different subjects to create new concepts or to reach a new level of understanding within the other subject. To apply this taxonomy to items it requires a different approach so that instead of describing the level of understanding a student possesses instead the item is classified based on what level of understanding the student requires to answer it.

Bloom's revised taxonomy and SOLO can be used to classify the items based on the cognitive level and order of thinking required by the students to obtain the correct answer; however, it does not address the construction of the item. The four major components that a MCQ is constructed around (stem, auxiliary information, answer option, distractors) can be classified in order to provide insight into how the item is constructed. The stem of the question provides the context of the question (the setting in which the question is asked), the content (the topic or concept being asked about), and should give some indication of the process the students should be undertaking to answer the item. Any auxiliary information can help add information to the context and the content of the item, and usually it will clearly highlight the process the students are expected to undertake. The presentation of the stem and the auxiliary information, as well as the options can also be used to describe how an item is constructed. This might mean that the question is more visually based as it includes diagrams, it might be presented as an equation that needs to be solved or require comprehension to make sense of a large amount of text. All this information can be used to describe the type of item that is being asked of the students, which is another way of potentially grouping items.

The expected difficulty of the item can also be used as a way of classifying MCQs. This can be informed based on knowledge of the topic, previous results of a related item, or the process that the students are expected to undertake to obtain the answer. It could also consider factors such as the number of steps the item requires (e.g. the number of equations they need to use, or how many diagrams they might need to draw), and the number of times the students are expected to have encountered similar items before, as this could be used as a way to predict the cognitive load placed on the students.[216] This is because the students would be expected to perform better on an item that they have seen and practised previously than an item that they have never seen before. However, it is also possible that an item that they have practised before but containing a new inclusion might be more difficult if the students are unable to account for the new inclusion.

It is also possible to use potential issues within an item as a form of classification. This is not ideal, as it would be preferable that items had no potential issues; however, if a new or existing item is a concern it is possible to highlight that when evaluating the items. These items would then be thoroughly analysed after the assessment has taken place to ensure that any of the potential issues did not cause problems. These might be items that include a known construction flaw (e.g. which of the following, all of the above, none of the above), potential for cueing (e.g. a relatively long answer option), or potential for misunderstanding/confusion (e.g. a comprehension-heavy mathematics question). [44,77,133,217] It is also possible that the item is a concern for other reasons, such as due to issues an item has previously displayed, or if there is concern over whether the students will behave the way they are expected. No matter the reasoning, identifying which items may be problematic before the assessment takes place could be a useful strategy for classifying items.

1.5 The Construction of Multiple-Choice Assessment

1.5.1 Constructing Multiple-Choice Questions

Many of the concerns surrounding the MCQ format are alleviated if the items are well constructed, as this will help prevent construct-irrelevant variance and increase the validity and reliability of the assessment.[34,43,45,75,134] However, many assessors are not aware of the ways in which MCQs can be constructed to improve these aspects of the assessments. There are two key aspects of MCQs that need to be considered when constructing an item: the stem and the distractors. Both have large influences on how the students approach and answer the questions, and thus both need to be seriously considered when writing questions. However, even before the stem and the distractors are explicitly generated, the content that is being assessed needs to be decided. Even at this stage in assessment writing there are several considerations that assessors need to be aware of to ensure that they create items that will give valid and reliable results. There are a number of guides that can be followed to ensure that any MCQs generated are free of the most common issues, but the key guidelines are summarised in the sections that follow.[38,43,44,46,75,218-222]

1.5.2 Assessment Content

The content that is assessed in a MCQ assessment needs to be carefully considered to avoid causing problems with student interpretations and answers to questions. Each item should focus on one specific concept, to ensure that the students are only being assessed on one aspect of the course at a time. It also ensures that the information being provided to the assessors can be easily interpreted, as if the students consistently answer an item containing multiple concepts incorrectly, it is unclear which concept is causing the issue for the students. There is also a concern that MCQs encourage more trivial content being assessed and more 'trick' questions; however, this is not the fault of the format but the fault of the assessors. Obviously trivial content and 'trick' questions should not be used within assessments, because not only do they introduce construct-irrelevant variance, but they also cause the students to stop learning the course content and instead learn for the assessment. Thus, the temptation to include any of those sorts of items needs to be avoided. Any content that is opinion-based is not appropriate for a MCQ assessment, as it undermines one of the greatest strengths of the MCQ format in that it is a purely objective assessment. The most important aspect of the concepts being assessed, regardless of the format, is that it matches the expectations that are described within the course syllabus. As soon as an assessment addresses concepts that are outside of the syllabus it means two things: firstly, that the performance of a student in the assessment no longer matches how that student is expected to perform within the course. Secondly, it means that

the students will likely no longer consider the assessment to be a fair representation of their ability in the course, as it is assessing them on topics that are not described as a part of the course. This then introduces more construct-irrelevant variance into the assessment, as when students consider an assessment to be unfair they will study for the assessment rather than the course content.

1.5.3 Multiple-Choice Question Format Considerations

The format of the MCQ assessment also needs to be decided before any of the stem or distractor construction begins, as that will inform how they need to be constructed. The different formats have all been discussed earlier (Section 1.4.3), and all of them can be used to construct valid assessments if they are used correctly. This means that the format should be chosen based on what the assessors believe to be the best way to assess the students on the content that they wish to assess. One important consideration is, regardless of the format, all of the items used within a MCQ assessment need to be independent of each other. This means that a student's response to one item should have no direct effect or influence on their response to any of the other items asked within the assessment. This is because highly dependent items can impact the validity of the assessment, as if the student makes a mistake on the first part they lose marks in all the successive parts that rely on a correct response to the first part. In addition, such dependent items cannot be marked for partial credit based on their working as they can be in constructed formats.

1.5.4 Stem Construction

The stem of an item needs to provide the context and the content of the question in one or two sentences, as excessive amounts of information can overburden the students and influence their performance.[45,77] This is to ensure that the approach the students should be taking is clear to them when they read the stem, and they should not need to look at the options presented to determine what is expected of them. This means that both the directions (e.g. calculate the value of x), and the central concept present within the stem (e.g. using Newton's second law of thermodynamics…) need to be clear. As a result of this, the language used when constructing the stem is extremely important, as it can have a significant influence on how the students interpret the question.[45] The stem should be constructed using only language that the students are expected to be able to comprehend as part of the course, and it should give the required information in the least amount of words possible. For example, in an assessment for a mathematics course, the stem should not use complicated English in its construction as that may disadvantage English as a second language students for a reason completely irrelevant to their mathematics ability. Thus, careful selection of the stem's construction and language will help to avoid introducing construct-irrelevant variance based on the student's vocabulary, and ensures that if a student misinterprets a question it is because they were not aware of something relevant to the course rather than due to unrelated comprehension or language skills. Similarly, it is important that the students can understand how they need to answer the item, which can be problematic if there is a lack of clarity within the stem itself. For example, if the stem is negatively worded and the response options are also negatively worded, the students may be more confused by the double negatives than the actual item, particularly if English is a second language.[133] To avoid any of those issues it is recommended that words like "not" and "except" are kept out of MCQs unless they have to be included, in which case they should be clearly highlighted within the stem, and the language used needs to be carefully considered so it does not unfairly disadvantage particular groups of students.[133,223]

One strategy to test the level of student understanding is to use different phrasing of concepts and give novel problems in the stem rather than repeating the language used in teaching and example

questions. The language used within the stem still needs to be understood by the students; however, by presenting the question in a novel way the students need to be able to approach the question differently using the same concept. This helps to identify the students who simply follow the process of answering the question rather than understand the concepts or the steps involved and is therefore a useful strategy within MCQ assessments to assess the students' learning within a course.

The wording and the presentation of the stem is dependent upon the MCQ format being used, as how the options are presented to the students is dependent upon the format being utilised. However, the most important aspect of any stem is that the students can understand what it is asking them to do. If the stem can impart the required information and guide the students without introducing any construct-irrelevant variance, then the stem has successfully completed its role within the item.

1.5.5 Distractor Generation

The types of distractor responses that are required is dependent upon the MCQ format being used, as each format requires slightly different styles of distractor. As the most common format is the single-best response format, this section will be related more towards distractors for that format. However, in most cases, the same considerations still apply to distractors within other formats, and thus everything discussed will be relevant to some extent to the other formats.

The most important aspect of any distractor is that it represents a plausible alternative to the correct option (the 'key').[44,224] This means that at first glance the students should not be able to eliminate any distractor as an incorrect option, and instead the students may need to seriously contemplate the content before determining their answer. The distractor options should be selected by students who hold misconceptions and incomplete knowledge about a specific concept or idea. If the distractors are not plausible alternatives, the students will always be able to select the correct option, not because they know the answer but rather because all of the other options are illogical and make no sense as answers to the question being asked. Depending on the question being asked of the students, sometimes plausible alternatives are easy to write, and sometimes they are difficult to write. This is because students are more likely to hold misconceptions on more difficult concepts and ideas, which makes writing distractors much easier, whereas on more fundamental concepts and ideas the students are more likely to understand the idea or concept being assessed. This is not an issue so long as the easy items are easy because of the content and not because of the distractors presented to the students. The number of distractors required in any item is dependent on the assessment, as it should be consistent throughout the assessment, but it tends to vary between two to four distractors per item on average. That being said it is not the quantity of distractors that is important but rather the quality of those distractors.[126] This is because if the student believes that an option is not a plausible alternative they will simply eliminate it from selection contention. This means that even if an item has twenty distractors if only two of them are plausible alternatives the item is essentially a three option item rather than a twenty-one option one.[188] So while four and five option assessments are the most commonly used, a two option assessment can be an equally or more effective assessment so long as the distractor is of high quality. Writing high quality plausible distractors is not easy, and this is where most of the time is spent when creating a MCQ assessment.

The two most common ways to approach distractor generation is either using an expert's knowledge within the course content or using common student errors. In both cases, the goal is to generate distractors that align with specific misconceptions the students are likely to select over the correct

option. Because of this, it is therefore also possible to use the distractors selected to obtain a gauge about what misconceptions are most common amongst the student cohort.[225] When writing these distractors it is important that the options are all kept consistent in how they are phrased and their length to ensure that there are no cues that the students can use to identify the correct option.[183] Other cues to avoid include absolute phrases (always, never, etc.), close association to the stem, grammatical errors, absurd or ridiculous options, an obvious answer option, and overlapping options.[45,46,77] The problem with all of these cues should be fairly self-explanatory, as they allow the students to either eliminate the distractor options or clearly identify the correct option. However, it can be hard when writing a MCQ to pick up on these issues, which highlights the importance of reviewing the questions both before and after they are used. It is also important to note that there are no steadfast rules to follow when constructing MCQs, as in some cases it might make sense to include some of these potential issues within the distractors. For example, on occasion a plausible distractor is one that lies in direct opposition to the correct option. Usually, if students see two options in direct opposition, they will immediately eliminate one based on their knowledge, the construction of the stem, or based on the other options. However, if the other options and the stem are written with this in mind, answering becomes entirely reliant on the students' own knowledge. What needs to be avoided is when one option is clearly wrong but is associated with one or two options that follow the same logic but are less obviously wrong. When that occurs students can eliminate several options from contention while potentially having very little idea about what the answer should be.

Another aspect that needs to be avoided within distractors is the use of negative phrasing, as well as using the options "all of the above" and "none of the above". The reason that negative phrasing should be avoided is because, similar to its use in the stem, it can cause unneeded confusion in the students that results in construct-irrelevant variance.[133] As a result of this, it is simply easier to avoid using it to avoid any problems it may cause. The reason that "all of the above" and "none of the above" are poor options is because students are able to obtain the correct answer without actually knowing the correct answer to the question.[217] For example in a five option MCQ where one of the options is "none of the above" if the student knows that none of the options are correct they therefore know that "none of the above" must be the correct option. However, the student obtaining the mark on this item tells the assessors nothing about the student's knowledge of question being asked, but only about their knowledge on the options presented to them. Similarly if "all of the above" is the correct answer to a question the student only needs to identify that two of the options are correct to know that the only plausible answer is "all of the above". The use of "all of the above" and "none of the above" as options shift the way that students approach the items, and they provide less information to the assessors as a result.

Every question should only ever have one truly correct answer, which sounds obvious but on occasion it is possible that distractors are generated that are technically true but lacking some key detail. Therefore, the format is called the single best response format, as the students are specifically instructed to choose the option that answers the question the best rather than the option that is true. Despite this, it can be unfair to the students to include options that are correct, but just less correct than another option. However, it is possible to check student understanding of concepts by including statements that are true as options that do not answer the question being asked of them. This is because the students will have to truly understand what the question is asking of them to select the correct answer rather than them selecting the first option that they know to be a true statement.

The last consideration surrounding distractor generation is how the distractors and the answer are presented within the item. This refers to both how the options are formatted, and the order in which the options are listed. In general, it is recommended that the item and in particular the options are formatted vertically rather than horizontally, meaning that they should be listed down the page. There is no explicit reason for this besides making it easy for the students to be able to identify each option, but the most important aspect of formatting is consistency, and it is easier to consistently format the options vertically rather than horizontally. Ideally, the options can be placed in logical or numerical order to avoid any cueing with option placement. However, if the options are unable to be ordered in this way, then the ordering of the options should be randomised to avoid influencing the assessment results accidentally, potentially due to influencers such as primacy effects (students preferring options that appear earlier) or cueing effects.[185,186,226-228]

1.5.6 Reviewing Multiple-Choice Questions

Once all the aforementioned steps have been completed, a MCQ item has been generated to be used within an assessment. Obviously the process of generating items needs to be repeated until the number of items desired is reached, which may correspond to the number of items required for the assessment or some other desired amount (e.g. enough to construct a large item bank). Due to the number of items that some assessors require the potential for automated item generation has been researched,[229,230] which makes the process of reviewing the items particularly important to ensure that no flaws have been introduced. Generating the assessment from the items can be done in several ways. In some assessments, the items may be placed based on their approximate difficulty so that students start with easier items and work their way to the harder ones; however, while this may be a constructive way of ordering items it does not make a difference to the students' results.[231] Other assessments may group items together based on the content that they are assessing, which may then be placed based on the order that the students were taught the content. The ordering of an assessment may also be completely random, either by randomisation or due to the assessors adding items in no particular order. Whatever the case, it is important that before the assessment is used that each of the items and the assessment task is reviewed to ensure that it will give the best possible measure of student ability.

The most important thing to avoid within any assessment is construct-irrelevant variance, which may appear due to issues with how the items are constructed. The flawed construction of an item could confuse knowledgeable students and may reward students who are unprepared for the assessment. These construction errors can range from simple spelling, punctuation, or grammar mistakes to much larger issues such as too much overlap between the stem and the correct option. All these issues have been discussed earlier when discussing the generation of the stem and the distractors. However, reviewing the items after they have been placed within an assessment provides a good opportunity to review the entire assessment to ensure that some of these issues did not slip into the items by accident. One way of checking the sensibility of an assessment is by having a colleague peer review the assessment.[47] This person does not need to understand the assessment and what it is asking of them, but rather they should be able to ensure that the stem and the options make logical sense.

After the assessment has taken place, it is important to re-review the assessment using the data obtained from the students' results. The results of the assessment can be used to inform the assessors of how each of the items performed, which can then be compared against the expectations of the assessors to determine if any items may have an underlying issue. The results can be further broken down past regular statistical analysis (i.e. mean, standard deviation, point

biserial coefficient) to provide more insight into what might be causing the issue. This in turn can be used to improve the items so that if they are ever re-used in future assessments, they will perform more closely to what was originally expected of them.

1.6 Analysing Responses to Multiple Choice Questions

1.6.1 Data Collection and Initial Impressions

Most commonly, MCQ assessments that are undertaken are marked using automated marking tools, and as such the information returned to the assessors is the option selected by the students and their mark or score for each item. In addition, any information that the students provide on their answer sheet is provided with those marks, which should be some form of identification for the students such as the student's name or identification number. This information is then used to assign the student their grade based on the number of correct options they selected within the assessment. This completes the basic purpose of that assessment, which is to measure the student's learning within a specific course and potentially assign them a ranking or grade based on that. However, the results of an assessment can provide more information about the students and their knowledge, both as individuals and as a student cohort, than what is learnt through the raw results of an assessment.

Whenever any MCQ assessment is analysed there are two assumptions that need to hold true in the case of all analytical techniques, as these assumptions underpin the function of all MCQ assessments.[232] The first assumption is that all of the items are independent of each other, as if the items are dependent upon each other it means that getting one answer wrong can have a chain effect on the results of other parts of the assessment. The main issue with this for MCQs is that it may be unclear where and how students made the initial mistake, and thus it is impossible to give the student any credit even if their working is correct but they started from the wrong place. The second assumption is that the assessment is built from a homogenous item set, which means that all the items should be assessing the students on content and concepts from the same course. This is important because it means that the ability measure (a measure of the performance level attained by a student that is representative of their skill within a particular content area) is a consistent measure of the same ability, rather than a mesh of a number of different abilities the students require in the assessment. If this were the case it would mean that the ability measure obtained from the assessment is indicative of the student's ability in a broad range of concepts rather than just the concepts relevant to the course. In general, to meet the assumption of a homogenous set of items, it means that every item asked of the students has to be relevant enough to the course that the student's final grade can be impacted based on their knowledge of it.

The initial impressions of the results can give the assessors a reasonable idea of the student's ability relative to the ability level of the other students within the course. However, how students are expected to score versus the reality of their score, as well as the significance of extreme scores, need to be carefully considered when comparing students. It is expected within assessments that as the questions become harder, more students will choose an incorrect option, but the nature of the expected drop-off changes how the results of the students are interpreted. The ideal case described by Guttman states that once the difficulty of the question surpasses the ability of the student then the student will obtain the incorrect answer on all of the questions past that point.[233] In assessments ordered by difficulty, it means that patterns such as 11100 and 10000 (where 1 represents a correct answer, and 0 represents an incorrect answer) are expected results when analysing the student

marks. In reality that is rarely the case, and this can be attributed to one of two possibilities.[232] The first is that the ideal Guttman scale can never be achieved as it requires an assessment that in practice can never be made (i.e. requires no guessing and a perfectly homogenous item set), and thus every assessment is an approximation of the Guttman pattern. The other possibility is that student responses are not binary (i.e. they either obtain the correct or incorrect answer) but probabilistic (i.e. they have a probability of obtaining the correct and incorrect answers based on their ability). A probabilistic pattern also makes more sense within a MCQ format where regardless of the question being asked the students always have a chance of selecting the correct answer, as the students will often guess which option is the correct answer if they are not confident in their knowledge. Whichever option is expected by the assessors will impact upon the analysis technique that they will want to undertake, as some are based solely on the results obtained and others are based on probabilistic models.[232] The other consideration is how extreme scores are treated within the assessment analysis. An extreme score represents either the maximum or the minimum possible marks on an assessment, which occurs when a student obtains full marks or no marks, or if an item's correct option is always or never selected. The important note about both students and items that obtain extreme scores is that essentially no information is obtained about either of them. This is because if a student obtains an extreme score there is no way to know either the floor or the ceiling of their ability as their true ability may lie relatively close to or relatively far away from where the results they obtained indicates. Similarly, if an item is always either correctly or incorrectly answered, there is no way of knowing if the students had any difficulty or if they had no idea how to approach the question.[234] Generally, within MCQ assessments it is unlikely that the minimum extreme score will be obtained, as given enough items and students eventually a correct answer will be obtained by chance. However, a maximum extreme score is possible within MCQ assessments, and if it occurs assessors need to be aware of its significance when analysing the results.

1.6.2 Normal Distribution

To apply any statistical tools or methods of analysis it is important to understand the foundations of statistical analysis to ensure that the method is suitable and the analyst can have confidence in their result.[235,236] The normal model is consistently used in statistics,[237] and is appropriate for any distribution of data that is unimodal (contains a single most frequently appearing value) and roughly symmetric around the average result. The reason that the normal distribution is so important within statistics is due to the central limit theorem,[238] which states that:

> *"Given a large enough sample size the sampling distribution of the mean for any variable will approximate a normal distribution regardless of the variable's distribution within the population"*

The sampling distribution of the mean refers to a histogram that is generated using the mean values from multiple sample populations as individual data points. The theorem also states that it does not matter what sort of probability distribution the variable follows within the sample population (e.g. normal, left-skewed, right-skewed, uniform, etc.); so long as that population has a finite amount of variance the central limit theorem can be used to generate a normal model from the sample means. Many statistics require a "large enough sample size" which generally refers to an $n \geq 30$,[239] however that can change depending on the experiment and the initial distribution of the population.[240,241] Another requirement of the central limit theorem is that the variables measured are independent from each other, which means that the result of one variable cannot influence the result of another.

A normal distribution is often referred to as a "bell-shaped curve" based on the characteristic rising and falling of the graph around a single point. The mean (μ) and the standard deviation (σ) of the dataset act as parameters around which the normal model is generated. The mean is represented by the peak of the curve and the standard deviation represents the spread of the distribution around the mode, as can be seen in Figure 1 below.

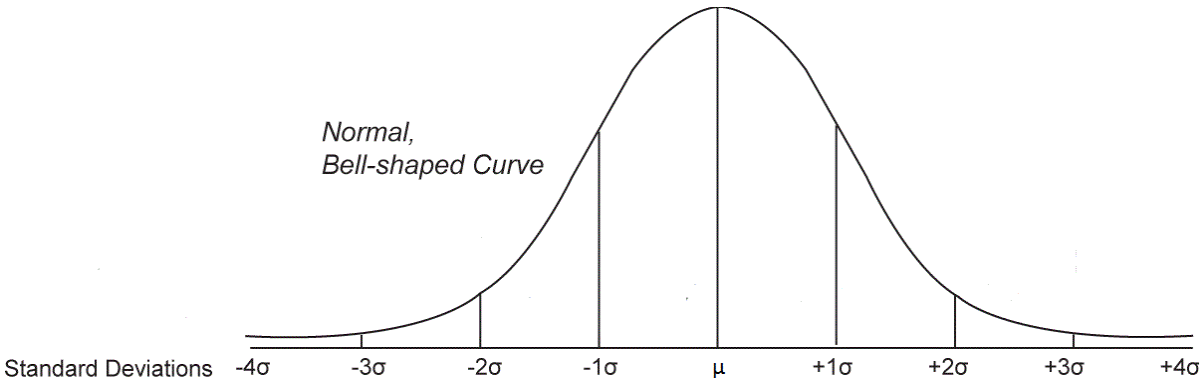

Figure 1: An example of a normal distribution (Adapted from "Normal distribution and scales.gif – Wikimedia Commons" by Jeremy Kemp; Public Domain)

Not every dataset follows a normal distribution, and even the data that fits the normal model are unlikely to perfectly fit the model. However, using the central limit theorem the approximation of a normal model can be justified, which is important as many statistical methods require the data to follow a nearly normal distribution in order to validate analytical techniques.[235,242] The normal model follows the distribution, represented in Equation 2:

$$p(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \qquad \textbf{Eqn. 2}$$

In Equation 2, *p(x)* represents the probability density function (normal curve) for a variable x whose distribution is centred on a mean value of μ_x and has a standard deviation of σ_x. This equation represents a perfect normal distribution and would generate the normal curve seen above in Figure 1.

A normal distribution only requires a mean and a standard deviation to be generated, and thus it can be expressed by the notation $N(\mu_x, \sigma_x^2)$. The square of the standard deviation (σ_x^2) represents the variance of the dataset, which is another way to describe the spread of the data around the mean. However, the true values of the population mean (μ_x) and standard deviation (σ_x) cannot be experimentally determined. This is because they represent the entire test population (*N*), and thus cannot be calculated based on the sample population that the dataset is generated from. Instead, they are estimated using the sample population (*n*), where the sample mean ($\bar{x}$) is used instead of the population mean. The formulas used to calculate the sample statistics are the same as those that would be used to calculate the population statistics, however instead they use experimental data, as represented within Equation 3.

$$\mu_x = \frac{\sum x_i}{N} \quad \approx \quad \bar{x} = \frac{\sum x_i}{n} \qquad \textbf{Eqn. 3}$$

Similarly, the standard deviation (σ_x) would be equated using the population mean and the entire population. However, instead it is estimated by the sample standard deviation (s_x) which makes use of the sample mean and the sample population, as shown within Equation 4.

$$\sigma_x = \sqrt{\frac{\sum(x_i - \mu_x)^2}{N}} \quad \approx \quad s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \qquad \textbf{Eqn. 4}$$

This is done because the sample population only represents a small subset of the entire population, and thus it is important to distinguish between the population and the sample being analysed. A new variance is also generated based on the sample standard deviation and is simply s_x^2.

It is not necessary to repeat the same experiment multiple times to generate a normal distribution through the use of the central limit theorem, and instead it can be generated from a single random sample.[243] This is because the mean of the sampling distribution ($\mu_{\bar{x}}$) is equal to the mean of the sample population ($\bar{x}$), and the distribution of the data can be determined from the sample standard deviation (s_x) and the sample size (n). This means that a new population mean ($\mu_{\bar{x}}$), and a new distribution descriptive (the standard error (SE ($\bar{x}$))) needs to be generated. The new population mean ($\mu_{\bar{x}}$) is the mean of the sample means, and the standard error ($SE(\bar{x})$) is dependent on the sample standard deviation and the sample size (n) as represented by Equation 5.

$$\bar{x} = N\left(\mu_{\bar{x}}, \left(SE(\bar{x})\right)^2\right) \quad ; \quad \mu_{\bar{x}} = \mu_x \quad , \quad SE(\bar{x}) = \frac{\sigma_x}{\sqrt{n}} \qquad \textbf{Eqn. 5}$$

As the sample size increases, the sample distribution will more closely approximate a normal distribution, as can be seen within Equation 5 and stated within the central limit theorem. This is important because a number of statistical tests require the data to follow an approximately normal distribution,[244,245] and thus the assumption of a nearly normal model is true in sample populations with large enough sample sizes. There are other simple methods that can give indications as to whether the data fits a normal distribution rather than justifying the fit using the central limit theorem. The simplest method of doing this is to plot a histogram of the observed data, as this can highlight any obvious trends and distributions within the data. Another method is to generate a Q-Q plot[246] of the data, which when plotted will appear as a straight line if the data follows a normal distribution, or will deviate from a straight line at points were the data does not follow a normal distribution.

The most common statistical test used that has an underlying assumption of normality is hypothesis testing, which is used to determine whether the statistical data accepts or rejects a belief about the data being analysed. The null hypothesis represents the belief that is held about the data, typically related to whether data fits within an existing model, and the alternative hypothesis represents the opposite of that belief. For example, if two sets of data are believed to belong to the same data set then the null hypothesis is that there is no significant difference between the two data sets. The alternative hypothesis in this example is that the two sets of data are significantly different from each other. To test the null hypothesis it is possible to apply one-sided and two-sided tests,[244,245] which estimate the probability that any given variable lies within the dataset. Both of these types of tests use a null hypothesis that assumes that the variable being assessed does lie within the expected range of the dataset, which is rejected when the probability of that being true is outside of a previously specified confidence interval (typically a 95% confidence interval is used). A null

hypothesis and an alternative hypothesis can never be accepted as true, only ever rejected, or retained based on the result of the statistical test. A two-sided test is used to determine whether some variable is significantly different from the mean in either direction (i.e. does it lie between $\mu \pm 3\sigma$). In contrast to this, a one-sided test is used to determine if the variable is significantly different from the mean in one direction (i.e. is it between $-\infty$ and $\mu+3\sigma$). A one-sided test is commonly used to determine if a variable is significantly different from another variable being tested without any regard for the opposite outcome. Both the one-sided and two-sided tests use the integral of the probability density function to determine the probability of the variable appearing within the specified range.

1.6.3 Z and T Statistics

The ability to construct a normal distribution allows for the use of hypothesis testing to determine the probability that a specific value lies within the dataset. While this can be done using the normal distribution generated from the observed data ($N(\mu_x, \sigma_x^2)$), it is possible to convert this distribution into a new model that has a mean of 0 and standard deviation of 1. The new model would therefore be represented by $N(0,1)$, and simplifies hypothesis testing. This model is generated by converting each observed data point (x) into its corresponding z-value, which is represented by Equation 6.

$$z = \frac{x - \mu_x}{\sigma_x} \qquad \textbf{Eqn. 6}$$

Z-scores makes the data easier to interpret, as now the confidence intervals can be described in terms of the number of units from zero instead of using the original values generated from the dataset. Using a two-sided test and the *z*-distribution, it is possible to generate the probability that a value would exist outside of $\mu_x \pm z\sigma_x$ by summating the area of the distribution that it corresponds to (see Figure 2). This can be used to show if values can be considered to be significantly different.

As the true population mean and standard deviation are not known, but are estimated based on the sample population, the *z*-value distribution will not give a perfect normal distribution, but instead it will form a distribution that will approach normality as the sample size increases. If the sample size is small (n<30) or the standard deviation of the population cannot be estimated, then it is not valid to generate a *z*-distribution. If this is the case and the population is approximately normal then a *t*-distribution can be generated instead.[247] The standard error ($SE(\bar{x})$) is used, rather than the standard deviation, as the t-distribution is highly dependent on the sample size (n), and thus this needs to be included within the equation, as shown by Equation 7.

$$t = \frac{\bar{x} - \mu_x}{SE(\bar{x})} \qquad \textbf{Eqn. 7}$$

As the sample size increases, the *t*-distribution has increased degrees of freedom (n-1) and as a result it begins to resemble the *z*-distribution more closely. The *t*-values generated represent the same values as a *z*-value, however *z*-values are more appropriate when larger sample sizes are used.[237] As a result of this, both the *z*- and *t*-distributions are used for the same statistical comparisons.

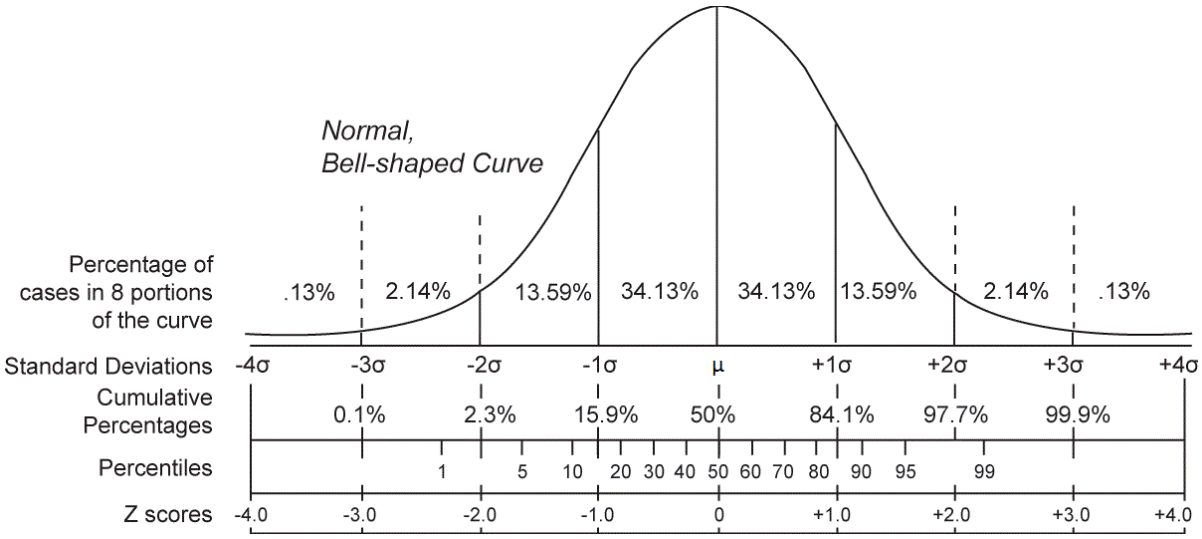

Figure 2: Normal distribution with *z*- and *t*-statistics, and the percentage that each section of the normal model corresponds to (Adapted from "Normal distribution and scales.gif – Wikimedia Commons" by Jeremy Kemp; Public Domain)

Using Figure 2 it is possible to construct confidence intervals for the dataset, which are then used as the criteria for retaining or rejecting the null hypothesis based on a reported *p*-value. For example, a two-sided 95% confidence lies roughly between μ-2σ and μ+2σ, or if *z*-/*t*-statistics are being used between -2 and +2 (a true 95% confidence interval is ±1.96σ). Similarly, a one-sided 95% confidence interval for a variable to be larger than the mean is between -∞ and μ+1.64σ. If the variable lies outside of the 95% confidence interval it is reported as having a *p*-value < 0.05, which is a statistically significant result in this example. Any value that lies within the specified range means that it is not statistically different from the mean value, and therefore the null hypothesis is retained. Any confidence interval can be constructed around a normal distribution using the proportions of the curve, however typically a 95% confidence interval is used for reasons discussed in Section 1.6.7.

Within every sample population there is a margin of error, which represents the deviation of the sample population from the entire population. This deviation from the population is measured around proportions within the data that represent different outcomes. The error is usually normally distributed around an observed proportion ($\hat{p}$), but this distribution is unreliable if the sample size is too small or if $\hat{p}$ represents an extremely likely or unlikely outcome. The mean of this distribution is equal to $\hat{p}$ and the standard error can be calculated using Equation 8.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \textbf{Eqn. 8}$$

While Equation 8 is stated to be accurate, providing $\hat{p}xn \geq 10$ and $(1-\hat{p})n \geq 10$ (where *n* – sample size), there are other methods that provide a more accurate calculation of the error.[248] This can be done using the Wilson score interval,[249] which generates upper and lower bounds of the observed proportion based on the confidence interval (*z*) desired, as represented by Equation 9.

$$p = \frac{\hat{p} + \frac{z^2}{2n}}{1 + \frac{z^2}{2n}} (\hat{p} + \frac{z}{1 + \frac{z^2}{n}} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4z^2}}) \quad \textbf{Eqn. 9}$$

While the previous tests described relate to testing how expected an observation was within a dataset, it is also possible to compare the difference between two different observed variables (x_1 and x_2) that are generated from independent samples. This can be used when comparing between two subgroups within the dataset to ensure that the assumption that they both follow the same distribution is true. For example, this can be used to test if there is a difference in how participants of different genders respond in a survey. This is tested using the null hypothesis that both values are equal, and thus when accounting for error $x_1 - x_2 = 0$. It is possible to compare the two variables assuming that the variables have equal variance (Student's *t*-test),[247] however it is recommended that unequal variances are used, and thus Welch's test should be used, as represented within Equation 10.[250,251]

$$t_{Welch} = \frac{x_1 - x_2}{\sqrt{SE(x_1)^2 + SE(x_2)^2}} \quad \textbf{Eqn. 10}$$

Based on this *t*-statistic, a *p*-value can be equated which can then be used to determine if the two values are within or outside of the specified confidence interval, and thus if the null hypothesis is retained or rejected. It is also possible to compare the means of more than two variables at a time using a one-way analysis of variance (ANOVA), which tests if there is any statistically significant differences between any of the means being looked at.[252,253]

1.6.4 Chi-Squared Statistics

Chi-squared (χ^2) distributions are composed of variables generated from the sum of the squares of *k* independent standard normal variables (i.e. variables that fit a normal distribution and do not influence each other), and hence they contain *k* degrees of freedom (the number of values that can vary when calculating statistics). The chi-squared statistic measures the degree of fit, and quantifies the extent of the deviation from the expected model.[254] This is extremely important when using the Rasch analysis, as it is expected that the observed data matches the Rasch model. It can also be used to compare the fit of the observed data to a hypothesised trend, which can be used to determine whether any subgroups within the dataset deviate significantly from the observed trend. A probability value can be generated from chi-squared values that can be used to perform hypothesis testing. The null hypothesis for any chi-squared test is that the observed variables match the expected variables, where the expected variables are calculated based on the model the data is expected to follow. If the null hypothesis is rejected, it implies that the sample population shows significant differences from the expected trend. Chi-squared values can be approximately converted to standard *z*-statistics using the Wilson-Hilferty transformation,[255] which is used within the Rasch model to more conveniently represent the results. A statistic following a chi-squared distribution (Y) with degrees of freedom (*k*) can be used to generate a standard normal *z*-value, given by *W*, as represented by Equation 11.

$$W(Y) = \frac{(\frac{Y}{n})^{\frac{1}{3}} - (1 - (\frac{1}{9})(\frac{2}{n}))}{\sqrt{(\frac{1}{9})(\frac{2}{n})}} \quad \textbf{Eqn. 11}$$

The chi-squared test for independence is used to determine if there is a significant difference between two or more subgroups within a dataset based on a trend that they are expected to be following.[256] This is done in order to determine if the subgroups within the dataset are independent of each other, and as such a contingency table is generated with the subgroups represented in the columns of the table and the outcomes in the rows of the table. The table needs to be generated twice, once with each cell calculated to reflect the expected number of data points for each row and column, and another time with the cells reflecting the observed data. The expected number of data points is based on the number of observations made, and the trend that the data is hypothesised to follow. The table is constructed with *r* rows and *c* columns, with each cell corresponding to a row *i* and a column *j*, with a total of *N* observations. The chi-squared statistic is then generated from the sum of the differences between the expected (*E*) and observed (*O*) variables with (*r*-1)(*c*-1) degrees of freedom, and from this a *p*-value can be obtained in order to test the null hypothesis, as represented by Equation 12.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \qquad \textbf{Eqn. 12}$$

This can be done to compare two or more groups of students who are expected to have the same probability of obtaining the correct answer in an assessment. In this case, the null hypothesis states that all the students have the same probability of obtaining the correct answer. If this null hypothesis is rejected, it means for some reason one group of students has a statistically significant higher chance of obtaining the correct answers in that particular assessment, which prompts further investigation into why that is the case.

1.6.5 Correlations

In statistics, it is common to structure models based on some number of variables that can be related by a mathematical function. If a mathematical model is generated that relates these variables, it is possible to use the model to predict the value of one variable using the function $f(x_i) = \hat{y}_i$ (where $\hat{y}_i$ is the predicted value of the i^{th} observed y value, y_i, corresponding to the i^{th} observed x value, x_i). From this, it is possible to generate the coefficient of determination (R^2), which quantifies the amount of variation predicated by the model, using Equation 13.

$$R^2 = 1 - \frac{\sigma^2_{residuals}}{\sigma^2_{total}} \quad ; \quad \sigma^2_{residuals} = \sum_i \frac{(y_i - \hat{y}_i)^2}{n}, \quad \sigma^2_{total} = \sum_i \frac{(y_i - \bar{y})^2}{n} \qquad \textbf{Eqn. 13}$$

This is used to describe the amount of variance accounted for within the model, where a value of 1 means that all of the variance is described by the model and a value of 0 means none of the variance is described by the model. Correlation therefore measures the strength of an association between two or more quantitative variables. In the case of two variables showing correlation (e.g. student ability and student results) there is an assumption of a linear relationship between two variables that follows the function described by Equation 14.

$$y_i = a + bx_i \quad ; \quad a = \bar{y} - b\bar{x}, \quad b = r_{xy}\frac{s_y}{s_x} \qquad \textbf{Eqn. 14}$$

Where s_x and s_y are the sample standard deviations of observed x and y variables, and r_{xy} is the sample correlation coefficient between the two variables that can be calculated using Equation 15.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad \textbf{Eqn. 15}$$

The standard error in the slope can also be equated based on this, as represented by Equation 16, and can be used in hypothesis tests between either the slope and a specific value, or between two different estimated slopes. In which case, the null hypothesis is that the value belongs to the model, or the two estimated slopes are the same as each other.

$$SE(b) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{(n-2)\sum(x_i - \bar{x})^2}} \quad \textbf{Eqn. 16}$$

1.6.6 Effect Size

To quantify the magnitude of the mean difference between two variables effect size can be used, which allows for the size of the statistical significance to be quantified. This is used complementarily to hypothesis testing, as whether the difference is statistically significant will be determined through other means. Effect size can be calculated using Cohen's *d*,[257,258] using the mean value for each data set divided by the pool standard deviation (which can be calculated through the use of Equation 18 and Equation 19), as shown by Equation 17.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} \quad \textbf{Eqn. 17}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad \textbf{Eqn. 18}$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_1)^2 \quad \textbf{Eqn. 19}$$

The results of Equation 17 will give a number between 0 and 1, where the larger the number the greater the difference is between the two means being analysed. Based on the size of Cohen's *d* the magnitude of the significance is defined by Table 1.[257]

Table 1: The classification of the level of the significance based on the size of Cohen's d

Small	0.20
Medium	0.50
Large	0.80

The main use of effect size is to complement other statistical methods, such that not only is it known whether something is statistically significant, but also how significant that result is. This can be used to evaluate the strength of a statistical claim, which is important when making major decisions based on the results of the statistical analysis.

1.6.7 Errors and the Bonferroni Correction

All the statistical tests described above are based on reporting the probability (*p*) that the observed data fits the null hypothesis. The null hypothesis is rejected when the *p*-value is less than a

predefined α-level, which reflects the confidence interval that is being used within the statistical test (e.g. a 95% confidence interval has an α-level of 0.05). The α-level is equal to the probability of a type I error occurring (the rejection of a true null hypothesis), which implies that a smaller α-level may help increase the accuracy of the test. However, decreasing the α-level increases the chances of a type II error (retaining a false null hypothesis), leading to a push-and-pull to optimise the α-level to account for these two error types. Therefore, the standard α-value of 0.05 is used in most statistical tests, including throughout this research, as it provides enough evidence to retain or reject the null hypothesis while also minimising the chances of both a type I and type II error. However, depending on the context of the hypothesis test, it might be important to change the α-level to reflect the level of confidence required to reject the null hypothesis. For example, in a medical context it is extremely important that the researchers are confident that the treatment is statistically different from the control, and hence they might use a smaller α-level to reflect that.

Another issue arises when multiple different hypothesis tests are undertaken on the same dataset, as each test increases the probability that a type I error occurs. If enough tests are undertaken it becomes increasingly likely that in at least one of the tests the null hypothesis is rejected. Without making any adjustments to account for this possibility, it is impossible to know if this is a significant result or if it is due to the increased chance of a type I error. The probability of a type I error is given by the family wise error rate, which is based on the α-level used and the number of statistical tests (k), as represented by Equation 20.

$$\bar{\alpha} = 1 - (1 - \alpha)^k \qquad \textbf{Eqn. 20}$$

Therefore, it is important not to utilise multiple statistical tests on the same dataset without consideration for the possible effects. However, if multiple tests are required of the same dataset it is possible to adjust the α-level to account for the increasing error chance.[259,260] This is done by applying the Bonferroni correction, which reduces the α-level in order to keep the probability of a type I error at or below the original α-level.[261,262] The Bonferroni correction is based on the number of statistical tests used (k) and the desired α-level, as represented by Equation 21.

$$\alpha_{corrected} = \frac{\alpha}{k} \qquad \textbf{Eqn. 21}$$

The corrected α-level is then used as the new significance level for the p-value, and thus to reject the null hypothesis the p-value must be less than the $\alpha_{\text{corrected}}$.

1.6.8 Factor Analysis

Within any assessment, it is important that the ability measure obtained from the students' results is the measure of only one ability and not the conglomeration of several different abilities. To ensure that the assessment can be considered to be testing only one ability measure, the results can be analysed using factor analysis. Factor analysis can determine how many underlying factors are involved in the generation of a single observed data point. The observed variable is then modelled as a linear combination of the potential factors identified plus an error term.[232] In assessment, it is assumed that the students' underlying ability is the most significant factor influencing assessment results, and any other factors influence the students at a level indistinguishable from random noise. Therefore, within assessments, factor analysis can be used to ensure that there is only one significant underlying factor influencing students' assessment results. If factor analysis finds a multitude of significant factors that contribute to the students' results it suggests that the

assessment is either not solely assessing the students' ability within the course, or the course is too broad to be defined by a single construct.

There are two broad types of factor analysis that can be used: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).[232] EFA is used to identify the relationship between different factors within a dataset where there are no assumptions made about those relationships prior to analysing the data. CFA tests hypothesised models against observed data to determine the amount each factor contributes to each data point, which is quantified by giving each factor a loading that represents how much it influences the observed outcome. The type of research being undertaken will determine the most appropriate type of factor analysis to be used, but in general EFA is used when nothing is known about the factors whereas CFA is used to test if the measures are consistent with current understanding. Generally, in the case of assessment, EFA is used because while ability should be the only factor that is responsible for the results of the students, that cannot be assumed to be true until it has been checked through analysis.

There are two common approaches within EFA that are used to identify how large the influence of the factors are on the results. The first approach is called common factor analysis and it seeks to identify the least number of factors possible to account for the common variance seen within the observations.[232] The other approach is principal component analysis (PCA), which seeks to explain all of the variance seen within the observations by continually adding factors to the calculations until all of the variance is explained.[232] Either approach is able to give results that can be used to inform the assessors of whether any factors outside of ability had an influence on the results of the assessment. However, what is important is the significance and the direction of each of the factors identified by these approaches. It is likely that the ability of the students will not be able to account for all of the variance seen within the assessment, as there will be other factors that influence the students' results, such as guessing. Because of this, it is important that the size of each factor is analysed to ensure they are above the expected noise level, and thus had a significant and consistent influence on the students' results. The other consideration that needs to be taken into account is whether the factor is positively or negatively influencing the students' results within the assessment. It is expected that the students' abilities have a positive influence on their results; however, other factors such as the students' tendency to guess may have a negative impact on overall results. Whatever the case, any factors that are found to significantly influence the students' results that are not directly related to their ability need to be thoroughly analysed to ensure that they are not impacting the validity of the assessment.

1.6.9 Classical Test Theory

Classical Test Theory (CTT) is a method for analysing the results of assessments that works by assuming that any score obtained by a student within an assessment is the combination of the score that their ability dictates they should receive (true score) and a random error factor.[232] The random error may either have a positive or negative effect on the true score of a student, but there is no correlation between the random errors for each student in an assessment. This means the error is truly random in the sense that it cannot be predicted or controlled for in any way, and thus the only way to measure the true score of the student is through the repetition of the assessment. This is because taking the average result over multiple sittings of the assessment will minimise the error, and eventually if enough results are gathered the true score of the student will be obtained through the student's average result. Of course, this does not work in reality, as the students cannot continually repeat the same assessment, as once the student has seen the assessment it will influence how they prepare and answer the assessment the next time, therefore changing the true

score that would be expected of the student. So, instead of this, students undertake multiple different assessments throughout a course, where the combination of all the assessment results is expected to provide a close approximation of the student's ability.

Classical Test Theory can easily be applied to any MCQ assessment, and with careful consideration can be applied to other assessment formats too. This is because the underlying assumptions of the model are relatively easy to meet. The first, and most important, is that the raw score obtained by the students within an assessment is the result of two components: the true score that the student should obtain, and random error that influences the true score.[232,263-267] It is expected that the random error is normally distributed, and thus over an infinite number of tests, the mean error should be 0 and the true score can be obtained. The standard deviation of the random errors gives the standard error of measurement, where the smaller the value the closer the raw scores are to the true score. Rather than have one student sit the same test multiple times in order to obtain enough samples to calculate an accurate standard error of measurement, it is possible to give the same test to a large student cohort and use the results generated from that one sampling in order to calculate the standard error. The size of the random error is dependent upon the reliability of the test. A reliable test will show very little deviations in the score of a student who re-sits the test, which is a result of the random error being quite small. An unreliable test will show large deviations in the student's score should the student re-sit the test, implying a large amount of random error within the test. Discrimination is related to reliability in that generally a poorly discriminating question does not give reliable results; however, it is possible that an item or assessment is poorly discriminating but still gives consistent results. CTT also assumes that all of the items carry the same weight within the assessment (i.e. the maximum mark for each item is the same), as when the items are compared to determine the reliability, it assumes that each item contributes equally to the student's ability measure. As is standard with almost all assessment analysis, the items are assumed to be independent of each other such that a student's result on one item will not impact their results on a different item. Since Classical Test Theory does not analyse the individual students, it assumes that the entire student cohort shares the same standard error. This assumption implies that the results of all the students who undertook the assessment all have the same random error around their true score. However, there is no reason to assume that the error of one student will match the error of another student, as it is more likely that error values shift depending on the result of the student. For example, it should be expected the students who obtain extreme results (0% or 100%) have higher errors as no information is gained about the limits of their performance, whereas students in-between these results clearly show their limitations and thus a better estimate of the error can be obtained.[263,265,267,268] These assumptions make Classical Test Theory the most approachable analysis technique. However, they also mean that some of the findings are hard to interpret, as a number of different considerations could be influencing the individual student outcomes and Classical Test Theory provides no direct evidence for the cause of the issue.[232]

Classical Test Theory focuses on how the items perform within an assessment, with the theory being if the items perform as they are expected to it will minimise the amount of random error in the students' results, meaning the results will be a more accurate reflection of the student's true score. As a result of this, Classical Test Theory is only able to evaluate each individual item and the entire assessment, but gives no information on individual student performance, as it only considers the entire student cohort. What this technique does evaluate is the item difficulty (the proportion of the cohort that obtained the correct answer), discrimination (how well the top quartile of students answers an item compared to the bottom quartile), and reliability (how well the performance of the item is correlated to the performance on the entire assessment), as well as the reliability of the

assessment as a whole (ensures that there is consistency throughout the assessment in the content being assessed) and the distribution of the results (ensures that the assessment is able to separate students based on their performance). This data can then be used to determine which items are causing issues for the students, and if these issues are a concern for the validity of the assessment. However, all these measures are dependent upon the results of the students. For example, if a student cohort has no issues with an item, it will have a low difficulty measure, but if a separate student cohort is asked the same item and they are unable to answer it then the difficulty will be much higher. This means that when considering what the results of the analysis imply about specific items, it is also important to consider the expectations the assessors have of the student cohort. If the assessment being analysed was used to determine how much the students know about a topic before it is taught within a course, then it should be expected that the students will find all of the items quite difficult. However, if the items being asked of the students are expected to be answered relatively easily, then a high difficulty should be cause for some amount of concern. The other issue with the analysis being dependent on the student cohort is that it can make it difficult to judge if an issue with an item is the fault of the item or the fault of the student cohort. This may mean that the same item has to be asked multiple times of different student cohorts before the assessors can be certain what the cause of the issue is and take action based on that.[232]

All these factors make Classical Test Theory a student-driven analytical technique that can only be used to identify if there are problems within the items or the assessment. This means that predictions about how an individual or group of students is expected to perform within the assessment are impossible to make using Classical Test Theory alone. Because of this and the fact that the results of the analysis are highly dependent on the student cohort it becomes difficult to determine if any issues detected within the analysis are due to the performance of the students or the items. This in turn makes it difficult to take any immediate action to improve assessments, as it is not immediately apparent what the best way to approach the issue is. This means that while Classical Test Theory is a very approachable method of analysis, there are some considerations that need to be accounted for before actions can be taken based on the results.

1.6.10 Use of Classical Test Theory

All of the calculations within Classical Test Theory are fairly straightforward, and only require the raw scores from any assessment undertaken in order to calculate the statistics.[266] Classical Test Theory can provide a basic overview of the assessment, giving information on both the individual items and the assessment itself. However, there are several limitations and assumptions behind Classical Test Theory that also need to be considered when using it to evaluate an assessment.

A result of the statistical methodology is that as more items are included within the assessment the assessment will become increasingly more reliable as a result of having a larger sample size,[269] which also makes small deviations appear as significant results.[269] This is because in statistics if a larger sample is used it means that the sample is expected to be a better representation of the entire population. As a result of this, the statistics generated from a larger population become more stable and less prone to random errors, which means that for Classical Test Theory the reliability of the assessment increases when more items are included. It is important to note that this is not always an issue, as it should be expected that increasing an assessment from five to twenty items increases the reliability, but at much larger sample sizes the reliability measures become slightly irrelevant as they are no longer an accurate representation of the reliability of the assessment.

Another consideration of Classical Test Theory is that all of the results produced are dependent on the student cohort that is being assessed,[266,270] and thus the results of the analysis will not necessarily hold true if a different student cohort is assessed using the same assessment. What this means is if an assessment is repeatedly reused each year the new student cohort may react differently to the items than the previous student cohort; however, that does not make the results from either cohort any less valid. While it is reasonable to have expectations for how the student cohort will react to an item based on previous years, without knowing for sure that the two student cohorts represent a similar population of students it cannot be taken for granted that that is true. It also means that the sample size is an important consideration when applying Classical Test Theory, as small cohorts will produce results that show large fluctuations in their measures.[271]

Classical Test Theory is heavily dependent upon looking at the random error within items and tests and determining their effect on the raw score. However, not all error within assessment can be attributed to random error and some of the differences in results can be attributed to systematic differences that occur when students sit an assessment on multiple occasions.[267] The changes in score between sittings may be a result of changes within the students themselves that might be the result of new learning and training within the area being assessed. These changes are more likely to help lower-scoring students improve than higher-scoring students, as they have more to gain in a pre- and post-test scenario. This means that the differences in the student scores cannot solely be attributed to random error, and in circumstances such as these, Classical Test Theory does not have any way to acknowledge that these changes are not the result of random error. This means any differences between the two assessments are the result of random error, which can affect the evaluation of item and assessment performance.

1.6.11 Rasch Analysis

Rasch analysis is a probabilistic model that was developed by Georg Rasch based on the idea that higher ability students have a higher probability of obtaining the correct answer than lower ability students.[272] Additionally, any student will have a higher probability of obtaining the correct answer to a less difficult item than a more difficult item. The principle of the Rasch model is:[272]

> *"A person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one"*

What this means is that students with higher abilities are expected to consistently perform better than students with lower abilities, and items with higher difficulties are harder to answer correctly than items with lower difficulties. There are four key assumptions that the Rasch model is built around, and these must be true for the Rasch model to be a valid method of analysis:[273,274]

1. Each person is characterised by an ability
2. Each item is characterised by a difficulty
3. Ability and difficulty can be expressed as numbers on a line
4. The difference between these numbers, and nothing else, can be used to predict the probability of observing any scored response

The first two assumptions describe the principle of unidimensionality within the Rasch model, which means that the performance of the student and the item is the result of only one attribute (ability and difficulty respectively).[275] One of the key aspects of measurement is unidimensionality, as it is

important that a single dimension or attribute can be measured.[276-278] This means that all of the items within an assessment need to be testing the same attribute, and the only factor affecting student results is their ability in the attribute that is being assessed. If this is not true then the assumption of unidimensionality is unsupported, as more than one attribute is significantly influencing a measure, and thus the Rasch model cannot be applied.

The third assumption discusses the scale that these attributes are measured on, as like any other form of measurement, the Rasch model requires units that quantify the strength of the attributes. The Rasch model expresses student ability and item difficulty along an interval logit scale,[273] which means that the distance between each logit value represents the same difference in student ability and item difficulty.

The logit scale represents a logarithmic transformation of the odds of success, which is based on the raw scores obtained within an assessment. The student's odds of success is based on their final score, where a raw score of 60% means that on any given question, without considering item difficulty, the student had a success-to-failure ratio of 60 to 40 (i.e. 60% chance of success, and 40% chance of failure). Similarly, the odds of answering an item correctly is based on the number of times it was correctly answered compared to the number of times that it was attempted, and that percentage is converted into a success-to-failure ratio, as represented by Equation 22.

$$logit\ (\log\ \text{odds}) = ln(\frac{success}{failure}) \qquad \textbf{Eqn. 22}$$

The transformation of raw scores onto the logit scale is critically important, as raw scores are able to order the students and the items based on their ability and difficulty but the differences in the scores do not have direct meaning.[279-281] For example, when considering the raw scores of the students, a difference of 10% means the same thing regardless of the final result of the student. However, a difference of 10% between high ability students (e.g. 85% to 95%) represents a larger ability gap than a 10% gap between average to low ability students (e.g. 45% to 55%). When converted into logits this ability gap becomes much more evident. Increasing from 45% to 55% might correspond to a change of 0.5 logits, whereas a change from 85% to 95% represents a change of 3 logits. This highlights why raw scores cannot be used as measures but can only be used to order the students and the items. In contrast to this, the logit scale is an interval scale,[276] which means that a difference of 0.5 logits represents the same gap in ability or difficulty regardless of where on the scale is being compared (e.g. -1.5 to -1.0 represents the same increase in ability as +2.5 to +3.0). The logit scale is used to calculate the probability of a student (n) correctly answering an item (i) based solely on the student's ability (β_n) and the item's difficulty (δ_i), fulfilling the fourth assumption of the Rasch model, as represented by Equation 23.

$$P_{n,i}(x = 1) = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \qquad \textbf{Eqn. 23}$$

Equation 23 is the foundation of the Rasch model,[272] as it directly relates the probability of the student obtaining the correct answer to the ability of the student and the difficulty of the item. When the student ability is equal to the item difficulty, the probability of obtaining the correct answer is 50%. On the logit scale the value of 0 is arbitrarily set as the average item difficulty, and the rest of the measures are given values relative to that zero point. Using the equated logits, the logit scale, and the probabilistic curve of the Rasch model it is possible to calculate the probability of

the students correctly answering any item within the assessment. This can then be used to determine how well the data fits the model, to evaluate the performance of the students and the items within the assessment and subsequently to evaluate the quality of the assessment. Originally, Rasch analysis also required the assessment items being analysed to not give partial credit, but models have been developed that allow for the use of Rasch analysis for items that allow partial credit.[282]

The probability of the students providing the correct answer follows a logistic curve, whereby the relative level of student ability and item difficulty determines the probability of the student obtaining the correct answer, as seen within Figure 3. The probabilistic nature of this model can be used to account for one of the largest concerns with a Guttman model, which is the chance that a student may obtain the correct answer through guessing. A probabilistic model can acknowledge that regardless of the ability level of the student, they always have some chance of obtaining the correct answer. Similarly, higher ability students still have a chance that they provide an incorrect answer to an easier item, which is also accounted for by a probabilistic model.

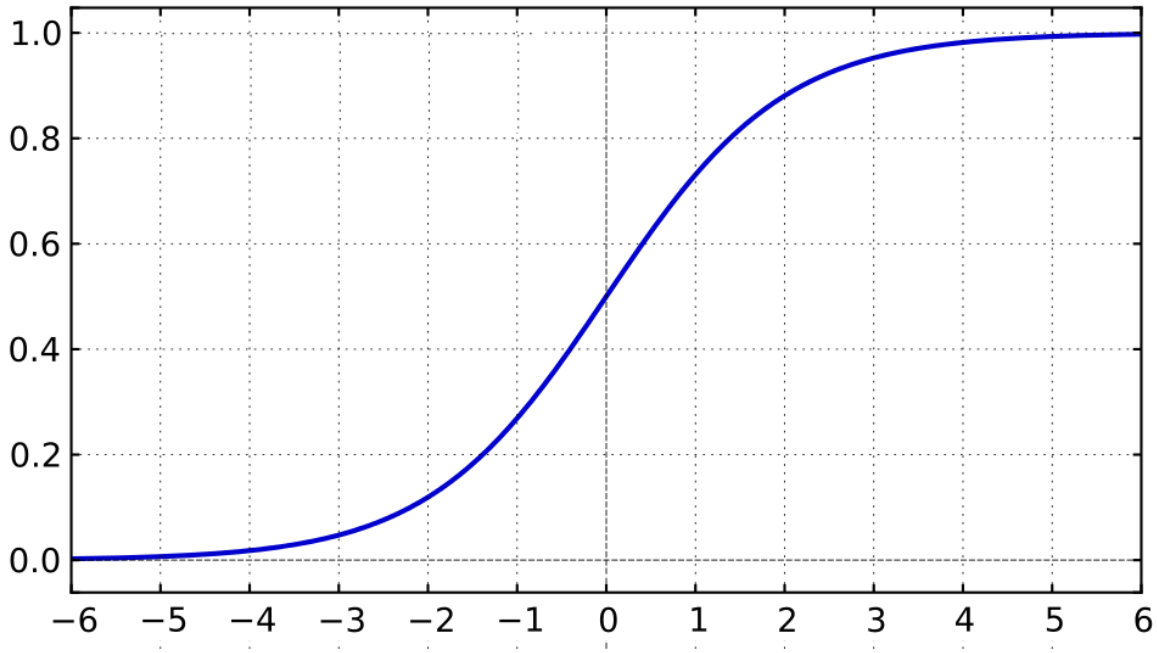

Figure 3: Example of a Logistic Curve which describes the Probability of a Student Obtaining the Correct Answer where the Student Ability Relative to the Item Difficulty (Student Ability – Item Difficulty) is Plotted on the x-axis and Probability is placed on the y-axis (Adapted from "Mplwp_logistic_function.svg – Wikimedia Commons" by Geek3; Creative Commons Attribution 3.0)

It can be seen within the figure that when student ability and item difficulty are equal the probability of the students obtaining the correct answer is equal to 50%. Also observable is that at the extreme ends of the curve it approaches 0% and 100% probability but it does not ever reach these values as they are asymptotes, which reflects the fact that no outcome is ever guaranteed within the Rasch model. Each individual item has its own unique logistic curve known as its Item Characteristic Curve (ICC); however, all these curves are expected to follow the trend illustrated above. The main difference between the ICCs, assuming that they match the expectations of the model, is that the x-axis will move left or right depending on the difficulty of the item. If the item is easier than other items within the assessment task then the x-axis will be shifted to the left, as students of lower

ability levels will have a higher probability of answering the item correctly than they would if the item was of average difficulty. If the item is harder than most of the items within the assessment task, then the x-axis will be moved to the right as the probability of the students answering the item correctly decreases for all but the highest ability students.

A key aspect of the Rasch model is that the student ability measures and the item difficulty measures can be considered to be completely independent of each other.[273,274,283] This means that through the use of the Rasch model "person-free" measures of item difficulty, and "item-free" measures of student ability can be produced,[273,284,285] which are critically important to any interval scale. Thus, the difficulty of the items is not dependent on the distribution of ability within the student population, and similarly the ability of the students is not dependent on the distribution of the difficulty within the items. Theoretically, because of this the same item difficulty and student ability should be obtained every time a student or item is measured, an important aspect of any form of measurement. An important note about this is that the measures themselves, while independent of each other, are placed on a relative logit scale. As such, it should not be expected that the exact same numbers for item difficulty and student ability are obtained every time that an item or a student is measured using the Rasch model. However, the interval difference between two difficulty and two ability measures are expected to remain the same, as their relative positions on the scale should not change (e.g. if items are apart by one logit on one scale then they are expected to still be apart by one logit on another scale even if their item difficulty measures have been shifted as a result of the new scale). This means that the Rasch model gives an invariant measurement of these values (i.e. the measurement of a variable lies on a standardised scale, e.g. distance, volume, etc.), something that is not obtained through other means of assessment analysis. This is significant as interval measurements are used as evidence to support or reject a theory or hypothesis, whereas in psychometrics the need for invariant measurement is often overlooked.

When conducting a Rasch analysis the data is expected conform to the assumptions of the model, and as such any results that do not fit the model are highlighted through the analysis.[273,286] Assuming that the model is correct, it means that there must be an issue with either the item or the student that misfits the model. It is therefore these items and students that should be further analysed to determine what may be causing any issues in the assessment. In the case of an item, it generally implies that there is an issue within the item that is either causing the students problems that are unrelated to the content being assessed, or the students are somehow taking advantage of the item. Either way, Rasch analysis can provide enough information that theories about what is causing the issue can be generated from only one sitting of the assessment. In the case of students causing an issue within the assessment, more information from Rasch analysis will help determine the issue, but it is likely due to either a high ability student choosing one of the incorrect options for an easy item, or a low ability student choosing the correct option for a difficult item.

It is possible to show the results of a Rasch analysis in a number of different ways, but the two most common representations are a Wright map and an item characteristic curve (ICC).[273,274] The Wright map displays the logit scale plotted with the ability of the students on one side of the line and the difficulty of the items on the other side of the line. This is an easy and effective way to visualise how well the assessment as a whole is measuring the students' ability, as ideally the average item difficulty and the average student ability should overlap to obtain the most information possible about the students from an assessment. An ICC is generated for each item used within an assessment and it can be used to show not only the probability of the students selecting the correct answer based on their ability, but also the probability that they select one of the distractor options.

Both plots are basic tools of Rasch analysis that can quickly and effectively show how well suited the assessment was for a student cohort, and how each of the individual items performed within the assessment.

Another important property of the Rasch model is that, as the measures of student ability and item difficulty are placed on an invariant scale, it is theoretically possible to compare the results of multiple assessments. As the scale is dependent on the assessment, there needs to be a way to link the different assessments that are going to be compared. This can either be done by using students who undertook both assessments, or by using items that were asked in both assessments. Through the use of 'anchoring', 'racking', or 'stacking' it is then possible to directly compare the measures from multiple assessments.[285,287] Anchoring is achieved by inputting the values of several student ability or item difficulty measures into the model before those values are estimated, effectively 'anchoring' those points within the model allowing for the other items and/or students to be fitted around those measures. Racking and stacking can be used when students have undertaken several assessments or when the items are used across several different student cohorts, respectively. This is achieved by 'racking' multiple assessments in the same row corresponding to the students who undertook all of the assessments, or by 'stacking' the responses to the same item within one column if it was asked on multiple occasions. However, it should be noted that this should only be used when it is expected that the assessments are testing the students on the same latent ability (the students' underlying ability within a particular content area that is attempted to be measured through their performance on assessment). Otherwise, comparing ability is completely invalid as there is no reason to expect the student to have the same measure for two different latent abilities (for more information see Section 2.5.5).

Like any model, there are assumptions and considerations that are not accounted for within the Rasch model that need to be addressed when it is used. There are two major assumptions that are not explicitly discussed within the model, which are that all of the items have equal discrimination and that guessing is not a significant factor in the probability of obtaining a correct answer.[273,288] Item discrimination is a measure of how well success on one item corresponds to success on the rest of the assessment. The assumption of equal item discrimination is important for the measures to be considered a part of an invariant scale, and as such, changes in item discrimination are not accounted for within the model. This assumption is difficult to justify, as it would be expected that more difficult items will show better discrimination than easier items, as commonly only higher ability students will be able to answer the more difficult item correctly. It is possible to calculate the item discrimination within the Rasch model after ability and difficulty have been determined; however, this value is merely used as a descriptive statistic. The assumption that guessing does not have a significant effect on the probability of success is similar in that there is no way to account for guessing when calculating the student's probability of success; however, it is different in that the assumption can be somewhat justified. Rather than assume that there is no guessing occurring within a MCQ assessment (which is obviously untrue), the Rasch model assumes that guessing is accounted for within the students' ability measure and random variation in the results. The theory being that given enough items, regardless of how much guessing occurs, the ranking of the students will become invariant, and any differences are accounted for by the variance allowed within the model. A final consideration is that because the Rasch model fits the data to the model, there is the potential for confirmation bias within the model itself and when reviewing the results of the analysis.[273] Simply because an item or student fits or does not fit the model does not determine immediately whether or not that item or student is problematic, but rather it highlights areas that should be looked into (see Section 2.5 for more information about applying the Rasch model).

1.6.12 Fitting the Data to the Rasch Model

In most statistical modelling, the model is structured based on the data, and any misfit is treated as the model deviating from the data. In contrast to this, the Rasch model emphasises the data's fit to the model, and as such any misfit observed is treated as the data not fitting the model.[289] This is because in order for ability and difficulty to be considered measures, they must fulfil the measurement criteria, one of which being that they need to fit a standardised model. Because of this, any misfit to the model is seen as evidence of poor construct validity (the degree to which an assessment measures what it is intended to), whereas in traditional statistics if the model does not fit the data it is simply seen as an inaccurate formulation of the model itself. This means that any ability or difficulty measure that substantially deviates from what is expected is a threat to the validity of the assessment and the measurement scale.

Misfit can be observed within the Rasch fit statistics (Section 2.5.3), and any students or items that are found to exhibit significant amounts of misfit need to be further analysed and potentially removed in order to improve construct validity. It is possible that a student may exhibit misfit if their answers to the items are not the result of the same underlying attribute as the other students. For example, it is expected that the students' ability trait is a measure of their competency within the course; however, it is possible that some of the students' scores may relate more closely to their "test wiseness"[290] (ability to guess answers based on the construction of the item) than their competency. This will result in these students having obtained several correct and incorrect answers that do not match the expectations of the Rasch model. Any assessment item that tests the students on anything other than the desired attribute may also exhibit misfit, as the students need to utilise a different set of skills to obtain the correct answer. However, it is also important to consider how wide or narrow the attribute being assessed is, as some courses can be quite broad in their topics, and it is possible that each topic needs to be considered as testing a different attribute. Flaws within the items themselves can also cause misfit, as they may help or hinder students in ways that are not reflective of the ability level of the students.

The Rasch model is therefore confirmatory in nature, as it is assumed that the data will fit the model, which is either justified or rejected based on the values of the fit statistics. Misfit is therefore a threat to the fit of the data to the model; however, despite this, misfitting data should rarely be removed from the analysis. This is because removing data introduces bias into the research. Data should only ever be removed when the objective is to obtain the best possible estimate of the ability and difficulty measures. If this is the case, then these estimates should not be used to justify any conclusions made in regards to the students or the items, as they are not true reflections of the measures that are obtained within the assessment.

One of the issues with gathering information from assessments in general is extreme scores, which correspond to the highest and lowest obtainable scores within an assessment. This is because they give no information about the ability of a student or the difficulty of an item.[291] If a student obtains full marks in an assessment then it is impossible to compare that student to any of the other students, as there is no way to know from the assessment the upper limit of that student's abilities. It is the same for students who obtain no marks in an assessment, as it is impossible to know how close their ability level is to the students who obtained one mark. Similarly, the difficulty of the items that are always answered either correctly or incorrectly are not comparable to the other items used within the assessment. This means that it is impossible for the Rasch model to assign measures to extreme scores, and instead the extreme score measures are estimated after the iteration stage has

been completed and all the other measures have been estimated. This is done by estimating the measure of a score that is slightly above the minimum and a value that is slightly below the maximum rather than using the minimum and the maximum scores themselves. This means that the estimated measures of extreme scores do not accurately reflect the measure of the student or item that they correspond to. Thus, it is common to omit these measures when calculating statistics reported in Rasch analysis.

1.6.13 Item Response Theory

Item response theory (IRT) is another probabilistic model that attempts to explain the probability of the students obtaining the correct answer by considering the student ability and item difficulty like the Rasch model. However, unlike the Rasch model, there are two additional parameters that IRT can include when calculating the probability of the students obtaining the correct answer.[288,292] The four different parameters are: student ability, item difficulty, item discrimination, and the chance of guessing, which can then be combined into three different models. Notably, the two additional parameters considered are factors that are not accounted for within the Rasch model. The inclusion of these factors as parameters is thought to improve the accuracy of the probability estimates that can be generated by IRT. Part of this increased accuracy is that IRT makes the model to fit the data, and thus the inclusion of more parameters can help to shape the model so that it explains more of the variance seen within the students' performance.

IRT has three assumptions associated with it: firstly, there is a unidimensional trait that defines a student's latent ability that can be measured on a scale; secondly, all of the items used within an assessment are independent of each other; and thirdly, the response of a student to an item can be modelled by an item response function (IRF). The IRF is used to determine the probability of the student obtaining the correct answer based on the student's ability (where a higher ability gives an increased probability) and the item parameters included within the model (1PL – item difficulty, 2PL – item difficulty and item discrimination, 3-PL – item difficulty, item discrimination, and guessing factor).[288] The item difficulty is used to determine the ability at which the students have a 50% chance of obtaining the correct answer (the same way it does within the Rasch model). The item discrimination determines how the rate of probability increases or decreases as ability changes, which is reflected within the slope of the IRF. For example, an extremely high discrimination will result in the students having an almost 0% chance of obtaining the correct answer until they reach a certain ability level when it will dramatically jump towards 100% chance. The chance due to guessing restricts the minimum probability of the students obtaining the correct answer to reflect the chance of any student correctly guessing the answer. This means that the shape of the logistic curve generated using the IRF is influenced by the number of parameters considered and their values.

Just like the Rasch model, IRT places its measures on a logit scale; however, unlike the Rasch model, these measures are not considered to be fundamentally item-distribution-free and person-distribution-free. This means that the measures cannot be considered to be a part of an invariant measurement scale.[293] This is due to the emphasis that IRT places on fitting the model to match the data obtained, and thus the model is built around the data rather than the data being fitted to the model. As a result of fitting the model to the data, there does tend to be less variance within IRT compared to other analysis techniques (depending on the number of parameters used). This means that while IRT models do not have invariant measurement they instead generate models that can explain most of the variance seen within any assessment.

1.6.14 Comparing Assessment Analysis Methods

While both Rasch analysis and IRT share many similarities in how the results of an assessment are analysed, CTT takes a completely different approach. This is because both Rasch and IRT use probabilistic models to determine the probability of a student's success, whereas CTT is focused on the performance of the items and assessment task. The other key difference is that all of the results obtained through CTT are entirely dependent on the student cohort that is being assessed, and thus while comparing scores using CTT is possible it cannot be justified mathematically that the values being compared are measuring the same trait. Comparing the results of two different CTT calculations assumes that the ability of the student cohorts are not significantly different if the item performance is being compared, or if the student performance is being compared it is being assumed that the difficulty of the items has not changed (whose measure is dependent on the performance of the student cohort). Thus, while it may be possible to draw potential conclusions about student cohorts by comparing the results from CTT, they cannot be confirmed until the other assumptions can be justified through some other means. In comparison, Rasch gives completely independent measures and IRT produces measures that can be justified to allow comparisons between the same latent ability trait. Another issue with CTT is that due to its statistical nature, assessments with more items are inherently more reliable due to there being more overall data, an issue that does not exist within probabilistic models. Despite this, it should be noted that CTT does have easier assumptions to meet than probabilistic models, and it is a more approachable method of analysis due to its simplicity. However, that simplicity comes with the price of obtaining less information about the assessment and the individual students.[283,294,295]

The similarities of the Rasch model and IRT are often noted, particularly since the 1PL IRT model is mathematically equivalent to the Rasch model. This means that the equation for the Rasch model and the 1PL IRT model are exactly the same, and thus often the two methods of analysis are confused and referred to as the same analysis.[273,283,288] The key difference between the two is how the models are applied when they are used to analyse assessment results. One of the defining traits of the Rasch model is that the data must fit the model, or the item is considered to be performing outside what is expected from the assessment. However, this is not true of IRT, as the model is built to optimise its fit to the observed data. This makes the Rasch model a confirmatory model, as the data must fit the model, whereas IRT is an exploratory model, as the parameters are generated to best explain the observed data. This leads to several significant underlying differences between the Rasch model and IRT that need to be considered when comparing the models. The first of which is that for invariant measurement, a confirmatory model is required, as the scale cannot be changed to suit the data being analysed. This means that whereas the parameters within the Rasch model can be considered to be measurements, the parameters within the IRT cannot be considered as measurements on an invariant scale. Therefore, while student ability and item difficulty can be thought of as fundamentally independent of the items used and the student cohort in the Rasch model, the same cannot be said for IRT measures.[293] The other comparison that is often drawn between the Rasch model and IRT is that the model generated by IRT generally gives a better fit to the observed data than the Rasch model. Again, this is due to the difference between a confirmatory model and an exploratory model. Within IRT the parameters are specifically generated to explain the observed data, whereas within the Rasch model the parameters must fit within the restrictions of the model or be defined as misfitting. The other consideration is that within 2PL and 3PL IRT models there are more parameters used to describe the data, and whenever more parameters are used it should be expected that a better model fit will be acquired. None of these considerations inherently makes one type of analysis more appropriate than the other, but they are factors that should be

considered by the assessors when they are choosing which model they want to use for their analysis.[273,288]

Regardless of the analysis used to determine the success of an assessment, it is important that at least some form of analysis is used to provide evidence that the assessment has adequately performed its role in assessing the students.[232,273,296-298] For some assessors this may be using CTT, as it is an easy and approachable method to obtain information about the performance of the items. Other assessors may prefer to use probabilistic analysis to obtain more information about the assessment so that it can be improved in the future. Whatever the case it is important that the assessment can be validated as a way of measuring the students' ability within a course, as if the validity of the assessment cannot be confirmed, then the calculated ability measures may be flawed. This is unfair to the students who have spent time, effort, and potentially money to succeed within an assessment only to be let down by poorly constructed items. This in turn will have flow-on effects to other aspects of the student's education such as their engagement and their approach to learning. It is important to remember that the results of any analysis are only ever as valid as the assessment, meaning if the assessment does not perform as it is expected to, there is not much information that can be gained from its analysis except that the assessment needs to be improved to more accurately reflect student ability.

1.7 Objectives of this Thesis

1.7.1 Research Questions

The goal of this research is to analyse the way in which MCQ assessments are used as an assessment format within first-year science courses at The University of Adelaide (specifically within Chemistry courses), to ensure that they are effectively measuring student ability. Doing so also provides an opportunity to explore how students answer MCQs, and the factors that can influence how they answer those items. Based on the data available, and questions surrounding student performance and assessment, six research questions were developed that are aimed to be addressed by this research.

> ***Research Question 1:***
> *Are the MCQ items used both previously and currently at The University of Adelaide in first-year Chemistry courses performing as they are expected to?*

The first step before exploring any deeper issues within an assessment is to first determine whether the items being used in that assessment are providing valid information about the students' knowledge and/or understanding. If an item is not performing how it is expected to within an assessment, then it is invalid to make claims about the performance of the students based on the information that the item is providing to the assessors. Thus, any problematic items that are present within an assessment need to be identified so that they can either be improved upon, or removed from the assessment, to ensure that the assessment is providing the assessors with information regarding student performance that is not tainted by poorly performing items.

> ***Research Question 2:***
> *Is there a significant difference in the performance of male and female students within MCQ assessments? If so, how can this be addressed?*

There is a high number of both male and female students enrolled in first-year Chemistry at The University of Adelaide, which allows for a comparison of the performance of male and female students within MCQ assessments undertaken. As there is no clear way to identify items or assessment tasks that contain some form of gender bias, this will involve the comparison of male and female performance at both the entire assessment task, and the individual items. The potential for other factors to be causing any differences in performance also need to be considered, such as the potential for student self-selection and the ability level of the students. Answering this question will not only help improve the items being used within the assessment, but it will also add to the growing knowledge regarding gender differences within assessments.

> ***Research Question 3:***
> *Do students show differences in their performance in MCQ assessments at different points in a semester? If so, how?*

There is a question of whether students change their behaviour toward assessment during the semester, either due to feedback, new information being presented, time pressures, or some other outside factor. Any change in their behaviour may affect their approach to learning and their application of knowledge, and thus may have an impact on their performance within an assessment. To research this possibility, the results of MCQ assessments that are undertaken twice at different times during the semester can be compared. Within the courses used in this research a MCQ assessment is undertaken during the semester that can then be resat during the final exam as a way to redeem marks. This allows for a comparison between the students' results during the semester compared to their results at the end of semester. While there are a multitude of possible factors that could influence the students, and be potential reasons for changes in academic performance, the identification of significant trends can be used to show if these factors tend to have a positive or a negative impact on student performance. There is also the possibility that there are no significant differences between student performances, which would then provide evidence of the reliability of MCQs. As the marks between the two assessments are redeemable and only the student's best result is used in their final grade, it is possible that students only sit the assessment once (either during the semester, or in the final exam). This may occur due to a student sitting the assessment during the semester and being happy with their performance, and as such they would rather spend time on other aspects than resitting an assessment they believe they've already performed to their highest ability. It is also possible that a student may simply rather not study for the assessment during the semester knowing that they have the option to take the assessment within the final exam when they need to be studying for the entire course anyway. The different approaches that the students have to the assessment gives the potential to find trends within student groups who only take the assessment once compared to the students who sit the assessment multiple times based on the comparison between their performances.

> ***Research Question 4:***
> *Do student cohorts show differences in performance over multiple years? If so, how?*

Often very similar assessments are used from year to year within courses that contain either no or very small changes to the items used. If a new student cohort undertakes the same assessment as the previous student cohort, then that assessment can give results that could be compared between the two student cohorts due to the overlap in the items being asked. In most cases (depending on the assessment being analysed) it would be expected that the results from the two separate student cohorts would not show statistically significant differences, as it could be assumed that the two

student cohorts have the same average ability. While this is a reasonable assumption to make, there is also a perceived sentiment that student cohorts perform worse every year, and thus based on this theory it would be expected that there is a statistically significant difference between two student cohorts from different years. The best way to test these theories is to compare the results of student cohorts over multiple years on the same piece of assessment, or an assessment that has only had insignificant changes to it over multiple years. MCQ assessments are uniquely suited to comparing student results over multiple years as the items asked are often only reworded or show no change at all between years, and when there is significant changes in the items being used between years, often enough items are reused such that a fair comparison can still be made. MCQ assessments are objectively marked, which means that any significant differences must be the result of differences between the cohorts and not a result of changes in the assessors. The data being utilised in this research contains a large number of overlapping items and there are multiple years' worth of data from previous student cohorts. This can be used to generate a clear picture of how the results of student cohorts are changing over time, and if there are any noticeable and significant trends.

> ***Research Question 5:***
> *Is it possible to compare student results across multiple courses from different disciplinary areas? If so, do students show similar performance across multiple courses?*

There is an underlying assumption within education that the best students in one course are likely to be the best students in another completely different course from a different disciplinary area. The students enrolled in first-year Biology courses at The University of Adelaide share a significant amount of overlap with the students enrolled first-year Chemistry courses, and as such there is an opportunity to test this assumption to a limited extent. This could potentially provide useful information regarding the transferability of student's latent abilities and study habits between courses. However, it cannot be explained within this research if that is true for every course, but merely whether it holds true within the assessment tasks analysed within this research.

> ***Research Question 6:***
> *What is the most appropriate way to analyse MCQs in order to provide an approachable methodology that can be used to improve assessments?*

This is an attempt to summarise the analytical techniques that will be used throughout this research into a helpful and methodical series of steps that can be used by others who wish to analyse their own MCQ assessments. This is arguably the most important aspect of this research, as the answers to the other research questions will not be applicable to every MCQ assessment undertaken due to differences within the courses and the student cohort. With this in mind, it is important to show others how they can replicate the results seen in this research with their own MCQ assessments; however, as deeper forms of assessment analysis are not often undertaken by assessors it is important that this provides an approachable methodology. This will hopefully have a twofold effect on assessment analysis. The first is that by providing an easy and approachable methodology for assessment analysis, assessors will be able to generate more information about the performance of their assessments in a time-efficient manner, which can be used to inform future decisions regarding their assessment. The second is that it might encourage more assessors to analyse their own MCQ assessments, which they may not normally do due to time constraints or lack of knowledge about assessment analysis. In both cases the MCQ assessments managed by these assessors should improve in their quality, and thus provide a better measure of student ability.

1.7.2 Project Objectives

The research questions are concerned with the outcomes of the research, but to achieve these outcomes there needs to be a deliberate approach to the research. This approach is stated in the project objectives, as they represent the steps that will be undertaken within the research to obtain the data required to answer the questions. The project objectives are heavily influenced by the research questions, as each one was developed based on finding ways in which the questions could be answered using the data available. To answer the six different research questions there are five project objectives that describe the different methods of analysis that will be used within this research.

> ***Project Objective 1:***
> *To assess the items used in MCQ assessments both currently and previously at the University of Adelaide in first-year Chemistry courses to determine whether those items are providing the assessors with information that reflects the ability of the students* ***[Research Question 1]***

The first objective ties into the first research question, and as such this objective is focused on analysing the MCQ assessments that are being used to assess first-year Chemistry students. This objective also provides feedback to The University of Adelaide regarding the quality of the assessments being used and will highlight any problematic items that are found within the assessment tasks. This objective will be undertaken by utilising both Classical Test Theory and Rasch analysis as ways of analysing student responses to each item, and using the data derived from this analysis, the quality of the assessment tasks and the individual items can be determined (Section 3.2 and Section 3.3).

> ***Project Objective 2:***
> *To analyse the construction of MCQ items utilised at The University of Adelaide in first-year Chemistry courses to develop a method of classification for MCQ items* ***[Research Question 6]***

To improve MCQs it is important to understand what is being asked of the student, how it is being asked of them, how the information is presented to the student, and the steps the student should be taking in order to answer the question. Gaining a better understanding of these factors can help to identify the intended outcome of any item and using analytical techniques it is possible to see how that intention has been interpreted by the students. Knowing these factors can also help to improve any problematic items, as the problematic factor could potentially be identified and changed within the item without having to construct an entirely new item that assesses the students on a similar, if not the same, concept. Knowing the factors in an item can also be used to improve the assessment as a whole, as it can help ensure that the assessment covers all the relevant ideas and concepts, and that the items are distributed throughout different topics in a way that reflects the importance of each topic. It could also be used as a way to determine the sorts of items that may need to be included in a different assessment format, either because MCQs are not effective at assessing that aspect of the course, or to avoid unnecessary overlap within assessments. The creation of a classification methodology will bring together the analysis of the all the items available within the data set, and current ideas and theories within the literature (Section 3.4.6). This will then be used to generate a number of categories and sub-categories that a MCQ can be described as based on its construction, content, process, and several other factors important in a MCQ. Ideally, this would act as the first step in analysing any MCQ assessment, as this breakdown can help any assessor classify

the assessment without having to consult each person that generated an item within the assessment.

> ***Project Objective 3:***
> *To compare the performance of male and female students in first year Chemistry MCQ assessments at The University of Adelaide to ensure that any difference in performance is a result of a difference in ability and not due to factors within individual items that influence student performance based on their gender* ***[Research Question 2]***

When comparing between male and female student results within MCQ assessments it is important to consider whether any differences are a result of the format and the items used, or if they are due to differences within the ability of the male and female student cohorts. Thus, this objective is focused on firstly determining if there are any differences in how male and female students perform, and based on that result whether any differences observed are the fault of the items, the assessment format, or the student cohort. How this is done is dependent upon the methodology being used, as CTT must assume that male and female students are expected to have an equal probability of answering the item correctly whereas Rasch analysis does not need to make that assumption. Whether these differences in methodologies change the results seen is something that needs to be considered during the analysis to ensure that this objective is completed as accurately as possible.

> ***Project Objective 4:***
> *To compare item and student performance within first year Chemistry assessments over the period of a semester, across multiple years, and against Biology courses using MCQ assessments undertaken at The University of Adelaide to determine if there are any differences in performance, and if they these changes are a result of the items or the students* **[Research Questions 3, 4, 5]**

To determine if there is a significant difference between MCQ assessment performance over the course of a semester, between years, or between courses, it is possible to use the results obtained to make a fair comparison. This can be done by tracking students and their MCQ assessment results within the same course, comparing the results of students across multiple courses, and comparing the ability of the student cohort between years. The best way to undertake this analysis is by using Rasch modelling, as it allows for the use of anchoring strategies (Section 2.6.3), but this could also be done utilising the assumptions of CTT if some compromises within the analysis are made. Both methodologies require some amount of overlap to make the comparison, which is why MCQ assessments are uniquely suited to this task as often they do not change significantly between years and they are marked objectively.

> ***Project Objective 5:***
> *To identify the most approachable and effective methods to analyse MCQs, and develop a process that can be used to improve any MCQ assessment* ***[Research Question 6]***

Even though classifying the items can provide enough information to inform many decisions about an assessment, the quality of the items remains unproven until they are used within an assessment. After an assessment has taken place, the performance of both the items and the students should be analysed to ensure that the assessment was a valid way to measure the ability of the students. However, reviewing the quality of an assessment can be difficult for inexperienced and time-deprived assessors who do not normally review assessments beyond the most basic measures. In

order to provide these assessors a methodology that they can feel comfortable using to review their own assessments, the techniques used in this research will be compared to determine which of them gave the most effective and relevant results for evaluating an assessment. The reason for doing this is to encourage other institutions to review their own MCQ assessments so this research can have a broader impact on MCQ assessments in addition to the assessments analysed within this research.

1.7.3 Potential Impacts of this Thesis

Some of the questions posed within this research attempt to answer questions that align with previous research done on MCQ assessments, and others explore ideas that have never been previously published within the literature. The evaluation of individual assessments is research that has previously been undertaken, however at the very least this will help improve the assessments being analysed within this research. Any problematic MCQs that are found within this research will be further analysed in an attempt to improve the item for future versions of the assessment. The analysis of these items will also be used in an attempt to determine what makes them function the way they do, which can then be used to identify factors that cause consistent issues within MCQs.

Research question 3, 4, and 5 all relate to student performance within MCQ assessments, and how performance changes between courses, cohorts, and over time, which is a concept that is often discussed but not often approached in this way. All three of these questions have common theories surrounding the expectation of student performance, and why student performance may be different. However, they have never been answered using anchoring techniques and quantitative analysis. If these questions can be answered in this research, the answers can be used to help inform decisions regarding the use of MCQ assessments, for example:

- Knowing whether students perform consistently between courses can help assessors and students be realistic about their goals and achievements within any particular course
- Determining if there is any significant trend in how the ability of student cohorts change between years can be used to inform the direction of future teaching and assessment to account for the observed trends
- Knowing whether or not students will perform differently at different points in the semester can help assessors determine the accuracy of results obtained throughout the semester, and potentially help decide when assessments should be taking place in order to obtain the most accurate measure of student ability

The most important aspect of this research is to provide other assessors with a process for analysing their own assessments. Thus, hopefully, the improvement of the assessments within this research can be replicated by others at different institutions based on the process presented. Not only will this provide assessors looking for a way to analyse their MCQ assessments with an optimised process, but it will also encourage assessors with no previous experience in assessment analysis to attempt to analyse their own assessments. Ideally, this research highlights that MCQ assessments are not merely a simple assessment method, but rather an educational tool that can be created and adjusted to fit into any educational objective so long as the assessors put in the time and effort required.

Chapter 2. Methodology

2.1 Data Collection

2.1.1 Ethics Approval

Approval to use the multiple-choice question (MCQ) assessments for the purpose of this research was given by The University of Adelaide Human Research Ethics Committee on the 2nd of November 2017 (Ethics Approval Number: H-2017-210). The students were not notified that the results of the MCQ assessments were going to be used for research. However, this was deemed by the Human Research Ethics Committee (HREC) not to be an issue as the students were required to sit the assessments as a part of the course, and because all the data was de-identified to ensure student privacy.

2.1.2 First-Year Multiple-Choice Question Assessments at The University of Adelaide

This research was undertaken at The University of Adelaide which offers four different first-year undergraduate Chemistry courses (Foundations of Chemistry IA, Foundations of Chemistry IB, Chemistry IA, and Chemistry IB). All these courses were used in this research, as well as some results from one of the first-year Biology courses (Molecules, Genes, and Cells) offered at The University of Adelaide. Within all first-year Chemistry courses there are two MCQ assessments undertaken during the semester, both of which are redeemable within a section in the final exam where often the same (or similar) MCQs are asked of the students. The students are required to undertake the MCQ assessment at least once; however, they may choose whether they wish to do so during the semester, during the exam or on both occasions. Regardless of the decisions made by the students they are awarded the best mark they achieved regardless of when that mark was obtained, which means sitting the test on multiple occasions is risk-free for the students. The Chemistry courses also have a small section within the exam that has a new set of MCQs that are not redeemable that was included within this research. Any results from the replacement exam that was required of some students was not considered within this research as the student cohort that undertook those assessments was too small to give meaningful results.

The assessments used in these courses do not change much between years, which means that what students are assessed on in one year matches closely what students from a different year were assessed on. Each individual MCQ therefore likely has multiple records of student responses from the same semester, as there is the potential that each question will be asked of the students twice in one semester, and across multiple years. This large data bank informed the research questions (Section 1.7.1), as the data available set the boundaries for what could and could not be achieved within this research. This is because the data was collected independently of the research being undertaken, as these assessments were used within the courses as a graded assessment, and thus this research had no input into what data was collected.

The assessment tasks and items used in this research are not included within this thesis, as many of the items are still used within assessments, and thus publishing any of those items within this thesis risks the validity of future MCQ assessments undertaken at The University of Adelaide. Discussion of results that are relevant to The University of Adelaide to ensure high quality assessment tasks and items are intended to be had separately to this thesis to ensure that the validity of the assessments is protected while still improving them based off this research.

2.1.3 Student Cohorts

The data used within this research will utilise several different cohorts of students, where some crossover is expected between the cohorts. As previously discussed, the student cohorts are made up of students enrolled within first year Chemistry courses at The University of Adelaide. The data utilised will consist of student cohorts that undertook first year Chemistry between the years of 2012-2015, giving a large sample over multiple years to be used within this research. The reason that only these years was used rather than more current data was due to the time involved in processing the data to ensure student privacy and conform to ethics requirements. Each one of these student cohorts can range from 400 – 600 students, which varies depending on the course and the year of enrolment.

It is expected that there is a large amount of crossover between students enrolled in subsequent courses, for example Chemistry IA students are likely to enrol in Chemistry IB. This gives the potential to compare the assessment results from these two courses by linking the assessments using the students that enrolled within both courses. It should be noted that students are not required to enrol in either the subsequent or the preceding courses to undertake any of the courses. However, for students to progress to the next stage of Chemistry they are typically required to undertake both courses. It is also expected that there is smaller, but substantial, crossover between the students enrolled in Biology and Chemistry courses, which will allow for those courses to also be compared to each other.

The student cohorts are made up of a wide variety of students, all of whom have personal variables that may affect their results. It is possible to assign these variables to the students, as this information is given to the university as part of the student's enrolment. These variables include factors like age, gender, previous results, parent's education, and degree, which can then be used within the analysis of student performance to determine if any of them is a significant factor in influencing the results of the students either positively or negatively. Within this research gender specifically will be analysed to ensure that neither the assessment task nor the assessment items are having a significantly influencing effect on one gender more than the other. Gender was assigned by the students during the assessment tasks, who were given the options: 'male' and 'female'.

2.2 Data Preparation

2.2.1 De-identification

There was no need to collect any raw data as the data collection was done as a graded assessment within each of the courses, meaning the raw data is the students' responses to the assessment tasks. However, as each of these assessment tasks were summative in nature it meant that individual students needed to be de-identified. Thus, to avoid any conflict of interest or invasion of privacy every student was given a new identification tag that could not be used to identify specific students but was used to track them between assessments and courses. The new tag allowed for the comparison and anchoring of student results without the risk of any of the students being able to be identified.

The student's background variables were also linked to the new identification tag to allow for the comparison of these factors to the students results. However, the level of detail provided within

each of these factors was carefully considered, as if enough information is given it is still possible to identify a person without a direct link to a name.

2.2.2 Data Received

The data received was the raw data from the assessments with the students' identification code being de-identified into a new code, and the names of the students removed. The students were de-identified in a manner not disclosed, such that there was no way of identifying the identity of the individual student. The raw data included the option selected by the student, their mark on each item, and their overall mark for the assessment.

2.2.3 Initial Analysis

The first step in the data analysis is to use exploratory statistics to determine some of the basic statistical information from that dataset, which was done using SPSS. This is used to identify basic information about the dataset such as its mean (and the mean's standard error), median, variance, standard deviation, minimum, maximum, and range. This was also used to test the dataset for normality both visually using histograms and Q-Q plots, and quantitatively using tests of normality. All the datasets used within this research have a high enough sample size that normality can be justified by the Central Limit Theorem. However, it is important to consider if there are any major deviations from that distribution within the datasets, as some of the statistical tests used within this research work under the assumption that the data follows a normal distribution.

2.3 Classical Test Theory

2.3.1 Difficulty and Discrimination

There are three different measures used to analyse individual items within Classical Test Theory (CTT); the first of which is the item difficulty (P). The item difficulty is the proportion of the student cohort who answered the item correctly. The number of students who obtained the correct answer (N_C) is divided by the number of students who sat the test (N) to obtain the item difficulty, as represented by Equation 24.

$$P = \frac{N_C}{N} \qquad \textbf{Eqn. 24}$$

The larger the item difficulty the easier the students found the item, as more of the students obtained the correct answer. The ideal difficulty level for any item is between 0.30 – 0.90.[270,299] Every test should contain at least a few easy items that test the students on ideas and concepts that they are expected to know, and some items that test deeper understanding that most students are unlikely to answer correctly. The reason that the items should not be too difficult or too easy is twofold. The first reason is that if all, or none, of the students obtain the correct answer then it is impossible to differentiate between the ability levels of the students, and thus the item provides essentially no information to the assessors about the student cohort. This is referred to as floor and ceiling effects.[291] The second reason is that to obtain the most information about a student cohort the item difficulty level should sit at 0.50, as it provides the highest probability of differentiating students based on their ability. However, as it is expected that all of the items within the assessment are testing the students on similar and related concepts it should be expected that there is some amount of inter-correlation between the student's results on one item and their results on a different one. As a result of this it would be expected that in a reliable and discriminating

assessment that having a difficulty level of 0.50 for all of the items would result in the same 50% of students obtaining the correct answer and the same 50% of the students obtaining the incorrect answer for all of the items.[300] Thus it would be impossible to differentiate between the students within those two groups. In order to avoid that situation it is best to incorporate items with variable difficulty levels in order to obtain enough information from the assessment to show a clear differentiation between the results of the students.[301] The reason that it is important to avoid making the items too hard (0.00 – 0.30) or too easy (0.90 – 1.00), even though this would still provide information on the students, is that within these ranges the information gained is not enough to justify the use of the items.[299] These variations affect the reliability and discrimination of the assessment, and thus in order to optimise the assessment it is best to avoid those difficulty levels.

The discrimination index (D) is the second measure which is used by CTT to determine how well an item can differentiate between students of differing ability. It does this by comparing the results of high ability students to low ability students, who are grouped based on the top quartile and the bottom quartile of students. The quartiles may be based either on the results of the assessment being analysed, or they may be based on the overall performance of the students throughout a course. The discrimination is calculated using the number of students in the top quartile who obtained the correct answer (N_H), the number of students in the bottom quartile who obtained the correct answer (N_L), and the number of students that are present within each cohort, as represented by Equation 25.

$$D = \frac{N_H - N_L}{N/4} \qquad \textbf{Eqn. 25}$$

Another way of considering the discrimination index is thinking of it as a comparison between how difficult the different student cohorts find the item, as described by Equation 26 and Equation 27.

$$D = \frac{N_H}{N/4} - \frac{N_L}{N/4} \qquad \textbf{Eqn. 26}$$

$$D = P_{Top\ Quartile} - P_{Bottom\ Quartile} \qquad \textbf{Eqn. 27}$$

While most commonly the discrimination index is calculated using quartiles it can be calculated using the top and bottom 25% - 33% of any student cohort, so long as the equation is changed to reflect whatever percentage is used.[302] The discrimination index is expected to have a value > 0.30, as the higher ability students are expected to provide the correct answer more often than lower ability students.[270,299] If the discrimination index gives a negative value it means that on that particular item the lower ability students have a higher probability of giving the correct answer than the higher ability students, which is a clear indication that a factor outside of student ability is affecting the test results. Values between 0 – 0.30 may be influenced by the difficulty of the item, as items that lie on the extreme ends (e.g. 0.00 – 0.30 or 0.90 – 1.00) will make it hard to see clear separation in the performance of the students. However, if the item difficulty does not lie at an extreme, then there is a concern that the item has difficulty differentiating between students of varying ability levels. This may then affect the reliability and the validity of the item as a way of measuring the student ability, as seemingly high ability students are being disadvantaged by the item in some way.

2.3.2 Point Biserial Coefficient

The third way in which CTT analyses the performance of individual items is using the point biserial coefficient, which measures the reliability of a single item. It does this by comparing the scores obtained on a single item to the total score obtained in the assessment, as it is expected that a reliable item will be consistent with the rest of the assessment. If the item is indeed consistent with the rest of the assessment it means that there should be a large correlation between the item score and the overall score. The point biserial coefficient (r_{pbi}) is calculated by comparing the average total score of students who correctly answered the item ($\bar{X}_1$) to the average total score of the students who incorrectly answered the question ($\bar{X}_0$), accounting for the standard deviation in the total scores (σ_x) and the proportion of the students who obtained the correct (P) and incorrect answer to the item (1 - P), as represented by Equation 28.

$$r_{pbi} = \frac{\bar{X}_1 - \bar{X}_0}{\sigma_x}\sqrt{P(1-P)} \qquad \textbf{Eqn. 28}$$

Equation 28 gives a value ranging from -1 to +1, where any negative values mean that the item has a negative correlation with the rest of the assessment, suggesting that what is being assessed within that individual item does not match what is being assessed in the rest of the assessment. Low positive values suggest that there is a small correlation between the item and the assessment, however it does not provide enough evidence to definitively state that what is being assessed within the assessment is consistent with what is being tested in the item. To be satisfied that the item fits with the rest of the assessment, the correlation should be $r_{pbi} \geq 0.20$; however higher values are better as they provide more definitive evidence that the item correlation is not simply due to chance.[270,303] Similar to the discrimination index, it is important to consider the difficulty level of the item being analysed before making definitive statements about what the level of correlation suggests about the item. Items with difficulty levels in the extreme ranges can result in unexpected correlation results that may or may not be reflective of how that item fits into the overall assessment. Thus, it is unfair to judge those items without first considering the factors that may be causing that correlation that are unrelated to the assessment.

The difficulty level, discrimination index, and the point biserial coefficient are the three statistics that are used within CTT to analyse individual multiple-choice questions. When used together they can determine if the item is performing as it is expected to, both in terms of the expectations placed on an individual item, as well as its performance within the overall assessment. However, it is also possible to evaluate the assessment to ensure that the assessment itself is reliable and discriminates between high and low ability level students.

2.3.3 Kuder-Richardson Reliability Index

All the other methods discussed are calculated for each individual item asked within an assessment; however, CTT also has two different ways of evaluating the assessment task. The first way to evaluate the assessment task is by calculating the Kuder-Richardson reliability index (r_{test}) of the assessment. This measures the internal consistency of the assessment, to ensure that the assessment is constructed using items that assess the same material. If the assessment is consistent with its material, then the students are expected to also show a level of consistency in their ability to answer the items presented to them. This can be calculated using the number of items within the assessment (K), the difficulty level (P) on each item (i), and the standard deviation of the total score (σ_x), as represented by Equation 29.

$$r_{test} = \frac{K}{K-1}\left(1 - \frac{\sum P_i(1-P_i)}{\sigma_x^2}\right) \quad \textbf{Eqn. 29}$$

Equation 29 is known as KR-20.[304] In the cases that the assessment being analysed is not marked as either correct or incorrect, potentially due to awarding partial marks, Cronbach's alpha[305] needs to be used in order to calculate the reliability index. In general, an $r_{test} > 0.70$ is reliable enough for the purposes of group measurement, and an $r_{test} > 0.80$ is reliable enough to assess individuals; however, higher reliabilities are always better.[270,299,306] If the reliability index is lower than these values then it is important to consider whether this is expected or not. Some assessments are constructed such that they test a wide variety of material, and thus if the assessment is broad enough it might be reasonable to suggest that the lack of correlation between the items and the assessment is expected. If a correlation is expected, and none is seen, then it is important to look at the individual item analysis to diagnose the issue. Items with poor discrimination and point-biserial correlations are usually the cause of any issues, as these items are often not consistent with the overall assessment and thus can affect the reliability index.

2.3.4 Ferguson's Delta

The discriminatory power of the assessment is measured by Ferguson's delta (δ),[303,307] which analyses the distribution of student scores over the possible range of scores. It is expected that the student score distribution should follow a roughly normal distribution, which would provide a broad range of scores. This is because it is expected that there are a few students with exceptional high and low ability, but many of the students are clustered towards the centre of the distribution in a well-constructed assessment. To equate Ferguson's delta the total number of students (*N*), the total number of assessment items (K), and f_i (number of students whose total score is equal to *i*) is used, as represented by Equation 30.

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - (N^2/(K\ +1))} \quad \textbf{Eqn. 30}$$

Equation 30 gives values from 0 to 1, where 0 represents no discrimination at all and 1 is the maximum discrimination possible. Generally a $\delta > 0.90$ shows that the assessment has a large enough discrimination between students to consider it a valid assessment.[299] If an assessment gives a lower δ than this it does not necessarily mean that the assessment is not a valid way of assessing the students, but rather the assessment itself needs to be considered for what should be expected of it. In some circumstances, it might be expected that the students are obtaining very similar scores, as the assessment might be an easy assessment to ensure students know the required background information. Alternatively, the assessment might be designed to be extremely hard to measure how much knowledge the students have gained throughout the course. It is also important to consider the number of items being asked of the students, as the fewer the number of items, the harder it is to discriminate between the students' results. While Ferguson's delta can be a helpful measure in ensuring that the students are distributed over the range of possible scores, as it is possible to be influenced based on the assessment task itself Ferguson's delta should be used to support conclusions rather than justify them. Therefore, if an issue is observed within Ferguson's delta, and none of these considerations are expected to have any influence on the assessment, then individual item statistics should be analysed, as these can help identify which items might be causing an issue within the assessment and justify any conclusions being made.

This means that in total there are five different values that Classical Test Theory generates to describe the performance of induvial items and the assessment task, all of which listed below with the desired ranges of their values within Table 2.

Table 2: The desired values for each test statistic that can be generated through the use of Classical Test Theory and what aspect of the assessment that it relates to

Test Statistic	**Desired Values**	**Assessment Aspect**
Difficulty (P)	$0.30 \leq P \leq 0.90$	Individual Items
Discrimination (D)	$D \geq 0.30$	
Point Biserial Coefficient (r_{pbi})	$r_{pbi} \geq 0.20$	
Kuder-Richardson Reliability Index (r_{test})	$r_{test} \geq 0.70$	Assessment Task
Ferguson's Delta (δ)	$\delta \geq 0.90$	

2.4 The Basics of the Rasch Model

2.4.1 Generating Rasch Measures

All of the Rasch analysis done in this research was carried out using *Winsteps*,[308] a Rasch analysis software package which is designed to apply the two-facet (student and item) Rasch model. *Winsteps* can convert raw scores into logit measures and generate a logit scale from those measures. This can then be used to compare the observed results against what is expected based on the Rasch model. *Winsteps* also provides statistics (Section 2.5.3) that are used to evaluate the fit of the overall data to the model, as well as the performance of individual students and the items. Other analytical techniques (Section 2.6) can be applied within *Winsteps* to gather more information about why the students and the items perform the way they do.

Winsteps optimises the logit measures by estimating them from raw scores in two separate phases. Initially every person is estimated to have the same ability (B_n), and every item the same difficulty (D_i), before these estimates are revised using the PROX estimation algorithm.[309] The revisions are based on the performance of the students and the items, generated from their raw scores, and by assuming that the data follows an approximately normal distribution. Each iteration continues to estimate the ability and the difficulty based on the mean (μ), standard deviation (σ), observed raw score (R), and the maximum possible score (N) for each student (n) and item (i), as represented by Equation 31 for student ability estimates and Equation 32 for item difficulty estimates.

$$B_n = \mu_n + \sqrt{1 + \frac{\sigma_n^2}{2.9}} ln(\frac{R_n}{N_n - R_n}) \quad \textbf{Eqn. 31}$$

$$D_i = \mu_i - \sqrt{1 + \frac{\sigma_i^2}{2.9}} ln(\frac{R_i}{N_i - R_i}) \quad \textbf{Eqn. 32}$$

The iterations continue using the PROX estimation algorithm until the change in the estimate is either less than 0.5 logits, or a previously specified criterion is met. The second phase of the iterations uses joint maximum likelihood estimation (JMLE) in order to improve upon the estimated measures.[309] The current estimated measure of ability or difficulty is used in order to generate an expected score that is based on the expectations of the Rasch model. The expected score is compared to the observed score (raw score) to determine if the two scores are within an error range

of each other. The new estimate is then generated using the difference between the observed and the expected scores to adjust the current estimate, represented by Equation 33.

$$new\ estimate = current\ estimate + (\frac{observed\ score - expected\ score}{model\ standard\ error}) \quad \textbf{Eqn. 33}$$

All the new estimates are obtained simultaneously using JMLE, and the estimates are continually iterated upon until the convergence criteria is met (typically stated as when the largest logit change is ≤ 0.0001). This gives ability and difficulty measures that are independent from student and item distributions that can be used to compare the observed results to the expectations of the Rasch model.

Each estimate has its own standard error, which represents the variance of a distribution around the "true" theoretical value and measures the precision of the estimates. When estimating the standard error for an item or a student, the other items and students are treated as though their estimates represent their true values. This means that the reported standard errors do not account for the imprecision within other estimates. However, any imprecision because of this is usually far less than the actual reported standard error. To definitely state that two values are different from each other, and thus definitively order the ability and difficulty measures, the estimates need to be more than three standard errors away from each other.[273]

2.5 Rasch Statistical Methods

2.5.1 Separation and Reliability

It is important within any assessment that the results of the students are an accurate reflection of their ability within the area being assessed, and the two most basic measures of this are the reliability and the separation of the students and the items. The student reliability refers to how replicable the ordering of the student ability measures is if they sat a different assessment with parallel items. Similarly, item reliability represents how reproducible the item difficulty order is if the items were given to a different group of students that behave in a similar manner to the original student cohort. The reliability measures for the students and the items range from 0 to 1, where 0 indicates the observed hierarchy is randomly arranged and 1 indicates that the results will always match the observed hierarchy. Both student and item reliability require a large spread of results within the ability and difficulty measures to clearly demonstrate a hierarchy within the results. This can be affected by the sample size of both the items and the students, as more testing provides more information to calculate ability and difficulty measures more accurately. It is also possible to improve the reliability with a targeted assessment (in which the average item difficulty matches closely with the average student ability) in order to obtain the maximum amount of information from each of the students and items. Ideally student reliability is > 0.80 and item reliability is > 0.90 as these represent thresholds that still allow for variance within the assessment results while still being able to clearly define the relative ability and difficulty of the students and items respectively. The calculation of reliability requires comparing the observed variance and the true variance, as shown in Equation 34; however, as the true variance can never be known, but only approximated, the reliability can also only be approximated. The approximation of the true variance is based on an upper and lower bound set by the standard error of the model (which represents the randomness predicted by the Rasch model), and the real standard error observed (which represents the randomness within the data that is contradictory to the Rasch model) respectively. As contradictory

sources of randomness are removed and the two standard errors become closer, a better approximation of the true variance can be obtained.

$$Reliability = \frac{True\ Variance}{Observed\ Variance} = \frac{True\ Variance}{True\ Variance + Error\ Variance} \quad \textbf{Eqn. 34}$$

The separation of the students and the items refers to the spread of the measures that lies outside of the error range. This is considered as the number of statistically different strata that can be identified within the sample, which refers to the number of subsets that are observed within the sample. The student separation informs how many levels of performance can be consistently identified within the test, so a separation of 2 means that the test can accurately identify both high and low achieving students. The same logic applies to item separation. The separation index can have values from 0 to infinity, where the larger the number the more subsets that can be identified within the sample. Separation is calculated by comparing the "True" standard deviation to the error standard deviation, as represented by Equation 35.

$$Separation = \sqrt{\frac{Reliabilty}{1 - Reliability}} = \frac{True\ Standard\ Deviation}{Error\ Standard\ Deviation} \quad \textbf{Eqn. 35}$$

A high student separation (student separation > 2) is expected of an adequately constructed assessment, as a low student separation indicates a large amount of error and inconsistency. This implies that the assessment has trouble effectively identifying the high performing students from the low performing students.[309] However, it is important to consider the purpose of the assessment when looking at the separation as it may require a larger or smaller amount of separation depending on the purpose of the assessment. A high item separation (item separation > 3) is expected when a large student sample with a broad range of abilities is assessed, as a low item separation is indicative of small student sample sizes that cannot confirm the difficulty hierarchy. Both of these are tentative guidelines, as they are influenced by the assessment being undertaken, and the purpose of that assessment.[273]

2.5.2 Observed and Expected

Once the Rasch model has been generated for any sample, it is possible to compare the observed results to the expected results based on the predictions of the Rasch model. This is done by inputting the student ability and item difficulty directly into Equation 23 (Section 1.6.11) to calculate the probability of the student giving the correct answer. This can then be used to calculate the exact match observed% (OBS%) and the exact match expected% (EXP%), which can be compared against each other to determine if the observed data is more random or more predictable than the model. The OBS% is calculated based on how closely the observations made match the expectations calculated using the Rasch model. The EXP% uses the calculated expected values to determine how closely these match the expected outcome based on the Rasch model. It is important to remember that the Rasch model includes a level of unpredictability within its modelling, and thus the predicted expected values will contain some amount of randomness. This is what causes the difference between the predicted expected values and the expected outcome. These values are then used as a comparative tool to determine if the observed data shows more or less randomness than the model predicts. If OBS% > EXP% the observed data is more predictable than the model, and if OBS% < EXP% the observed data is more random than the model. Ideally OBS% = EXP%, as that means the data fits

the model's expectations perfectly; however, small differences between the two values are not a cause for concern.

It is also possible to calculate a correlation between the ability level of the students obtaining the correct answer and the item difficulty. This can be done for both the observed data, and the expected results of the Rasch model. This correlation is reported as a value between -1 and +1, where a negative correlation means that the lower ability students are obtaining better results on the item than the higher achieving students. It is expected that a positive correlation is observed within all the items and the correct answer option. It is also expected that the incorrect answer options should show negative correlations, as they are expected to be chosen by lower ability students more often than high ability students. This can quickly highlight whether the results of the items match the expectations of the Rasch model. It is also one of the best indicators for issues with data entry or the item key, as a negative correlation may be the result of human error. The difference between the observed and expected correlation works in tandem with the OBS% and EXP%, as it shows whether the data is more or less predictable than the Rasch model.

2.5.3 Infit, Outfit and Standardised Fit Statistics

While the standard error is used to report the precision of the estimates, fit statistics are used to determine the accuracy of the estimates. The two most commonly used fit statistics in Rasch analysis are the infit (inlier-fit) and outfit (outlier-fit) values.[310] The first step in obtaining the infit and outfit values is to compare the observed score (X_n) and the expected score (E_n) for the same data point. This is done by calculating a z-score for each data value (Equation 6, Section 1.6.3) using the standard deviation in the expected mean value (S_n) (Equation 4, Section 1.6.2) and the expected scored response for each data point. This *z*-score is termed the "standardised residual" (Z_n) for that data point. The infit and outfit are mean square values ($MnSq_{INFIT}$ and $MnSq_{OUTFIT}$) which are calculated using the standardised residuals from each data point and the total number of data points (N), as represented by Equation 36.

$$Outfit = MnSq_{OUTFIT} = \frac{1}{N}\sum_{n=1}^{N} Z_n^2 \ , Infit = MnSq_{INFIT} = \frac{\sum_{n=1}^{N}(X_n - E_n)^2}{\sum_{n=1}^{N} S_n} \quad \textbf{Eqn. 36}$$

The infit value is an information-weighted fit, which means that it is more sensitive to the pattern of responses around the difficulty measure of the item being analysed. This means that if an item has a difficulty measure of 1.00 then the infit value is more heavily influenced by the results of the students whose ability measure is close to 1.0. In contrast to this, the outfit value is not a weighted fit, and as a result, it is influenced evenly by every student response to an item regardless of the relative ability measure. This means that outfit values can be heavily influenced by outliers in the data, and as a consequence, highly unexpected results (such as a low ability student correctly answering a difficult item) can have a large impact on the outlier value. Both mean-square statistics show the size of the randomness (the amount of distortion) the data exhibits around the Rasch model.

Both the infit and the outfit values are expected to have a value of 1, where the amount under or over 1 represents the percentage of how much less or more variation is within the observations than the model. Values substantially above 1.0 are termed as displaying "underfit" and values substantially below 1.0 are termed as displaying "overfit". If the statistics show underfit then it means that the observed data is more random than the Rasch model accounts for, meaning there is

a large amount of inconsistency between the predicted and observed outcomes. Conversely, if the measures overfit it means the observations match so closely to the expectations of the model that the observed data shows less randomness than the model predicts. While the guidelines about what justifies underfit and overfit are different depending on the field of study, typically values between 0.70 and 1.30 are reasonable ranges for MCQs.[273] Within high stakes assessment, such as entrance exams to medical degrees, the ranges of 0.80 to 1.20 are used as it is important to ensure that there is no chance that the results are not purely reflective of the ability of the students.[273] In this research the ranges used within high stakes assessments was used (0.80 – 1.20) where any items found outside of these ranges were deemed to warrant further study. This was done due to the large amounts of data that was available to review, as it was possible to consider how often items appear outside of these ranges as a criterion for deeming an item flawed. Any item that had values outside of this range on only one occasion (assuming the item was utilised multiple times) was not considered flawed, as this was likely the result of random variation causing the item to appear outside the harsher guidelines that were employed. If an item consistently appeared outside of the harsher range, but within the regular range (0.70 – 1.30), it was considered to contain a minor flaw that still needed to be addressed in some manner; however, if an item consistently appeared outside of both the harsh and regular ranges then it was considered to be a majorly flawed item that either needed to be improved upon or removed from the assessment.

Outfit measures are less of a concern than infit measures, as the influence of outliers on the outfit value means that most significant outfit measures are a result of an outlier. This means that if the infit value fits within the expected ranges, but the outfit value shows a significant misfit it is likely an outlier causing the issue. In contrast to this, usually when there is a significant infit value the outfit value will also be significant. However, it is much harder to diagnose what is causing the misfit within an item when the issue is centred on the item's difficulty measure, and usually requires deeper analysis to identify the issue.

Standardised fit statistics, ZSTD, are t-tests that test whether the data fits the model perfectly and are reported as *z*-scores. They are used to evaluate whether any deviation from the model is within the expected amount of error or if it represents a statistically significant deviation from the model. If the data fits the model then the reported *z*-score should be 0, if the data are overly predictable it will give a value < 0, and if the data are more unpredictable than the model it will give a value > 0. As this is reported as a *z*-score it means that it has a standard deviation of ±1, and as such any ZSTD that is less than -2 or greater than +2 means the value is statistically different from the expected value. However, the ZSTD tests of the data in comparison to the Rasch model test whether the data perfectly fits the model, whereas the infit and outfit values test whether the data fits the model usefully. This means that the ZSTD values should only be used to determine if the deviation from the model is statistically different if the infit or outfit is also found to be statistically significant.[310]

2.5.4 Item Discrimination

Item discrimination is usually used within item-response theory (IRT) as the second parameter for determining student performance on an item,[232] but it is also possible to calculate item discrimination within the Rasch model. This is performed after the student ability and item difficulty have been estimated in a post-hoc analysis. Within the Rasch model, item discrimination is not used as a parameter, but used as a descriptive statistic. When calculated in a post-hoc analysis the item discrimination is calculated as if it were a third parameter within the Rasch model, but it does not affect the estimation of the other parameters as it would within an IRT analysis.[311] The item discrimination is also referred to as the item slope, and is considered to have a uniform value of 1

when the measures are initially estimated. The discrimination follows an asymmetric relationship with the fit statistics, where a low discrimination indicates a poor fit to the model due to unpredictability, and a high discrimination indicates that the observations are too predictable. The discrimination can have values from -∞ to +∞, but for useful measurement it is expected to lie between 0.50 and 2.[309] Any values lower than 0.5 means that the results are too unpredictable to be used, as they degrade the quality of the measures. Values larger than 2 do not degrade the measures, but they are likely the result of distortions within the assessment that usually indicate an issue within the items being used, either due to some bias or their ability to be 'gamed' by the higher ability students.[312] While the item discrimination is a useful descriptive statistic, it is expected to reflect the results already seen within the infit and outfit measures.

2.5.5 Dimensionality and Factor Analysis

It is possible that linear models are formed based on the relationship of more than two variables, but the Rasch model is based on the assumption that only two variables are responsible for students' results within assessments (student ability and item difficulty). Therefore, it is important to check the dimensionality of the observed data, as it ensures that the results can solely be attributed to student ability and item difficulty rather than some other factor. Factor analysis is based on describing the variability among observed variables in terms of a minimal number of unobserved factors.[313] The observed variables are then modelled as linear combinations of these factors with an error term; however, there are two main concerns with the use of factor analysis within assessment analysis. The first is that factor analysis does not provide information about which items and students define the underlying factors though the use of fit statistics (e.g. the factor may be the result of only one item). The second is that factor analysis is based on sample-dependent correlations, which makes reproducing the factor loadings (the amount each factor contributes to the model) on new datasets extremely difficult.[314-316]

Despite these concerns, it is important to confirm that there are only two major factors influencing student results to validate the Rasch model. Thus, a factor analysis is performed as part of Rasch model on the residuals (error between the observed data and the model) that remain after the variance explained by Rasch is removed. This factor analysis is used to identify any common source of variance that is shared amongst the data that is unmodeled or unexplained by the Rasch measures. The size of a factor (the amount of variance it accounts for) is reported for each contrast that is made (each factor being considered) as an eigenvalue ordered from the largest contrast to the smallest. The contrast of a factor needs to be above the noise level of the data before it can be considered an important factor that influences assessment results, within Rasch analysis the noise level is considered to be anything with an eigenvalue less than 2, and thus a factor needs to have a contrast of eigenvalue 2 or above to be considered to have a significant influence on the assessment. If there is a factor that is above the noise level then it implies that there are at least three factors that are influencing the results of the assessment, and thus the assumption of unidimensionality is flawed. However, this does not always mean that the results of the entire Rasch analysis is flawed, as this result may be due to misfitting items or students that are causing a new significant factor to appear. Therefore, if an assessment is expected to be unidimensional but it was found not to be then the items and the students should be analysed first to ensure that none of them are misfitting and influencing the dimensionality results. If no issues can be identified within the item or student analysis, and there is no obvious reason for the additional factors based on the results of factor analysis then the assessment may be flawed in some way and should be analysed to identify this issue. If there is the potential that the assessment may contain multiple additional factors (e.g. due to the presence of significantly different topics within the assessment) then that potential should be

explored through factor analysis by identifying which items and/or students are largely influenced by the additional factors and if that result can be rationalised. While there may be validity concerns for a Rasch analysis conducted on a non-unidimensional dataset there is the potential to justify the existence of additional factors; however, if that cannot be done then the results of the Rasch analysis should be carefully considered before they are acted upon.

2.6 Approaches to Data Analysis within the Rasch Model

2.6.1 Bias Analysis

Differential item functioning (DIF) and differential person functioning (DPF) are both examples of bias within assessments that can be identified using the Rasch model. DIF occurs when items are found to have significantly different difficulty measures for separate students or student subgroups. This can be used to identify items that have differing effects on a distinct student subgroup that might be based on the student's gender, age, or background.[317] DPF is similar except that it is the measure of student ability that changes in response to a particular item or group of items. This may occur when the students find one particular topic to be much easier/harder, if an item has construction issues, or if an item does not match the rest of the assessment.

There are two possible ways to measure DIF and DPF effects: the Mantel-Haenszel statistic[318] and the Rasch-Welch t-test.[250,309] The Mantel-Haenszel statistic normally requires complete datasets; however, while it is less accurate, it is possible to correct for an incomplete dataset.[309] The statistic is calculated by dividing the sample into classification groups (the subgroups being analysed) and then placing them into different strata depending on their measures (e.g. low, average, and high ability students). Usually, the statistic would classify the data based on the raw scores; however, using the estimated measures makes it possible to include observations that have missing data. A cross-tabulation is constructed within each of the classification group strata compared to the score response, and from this an odds-ratio is generated. A homogeneity chi-square test is conducted to test the null hypothesis that the subgroups come from the same population. The Mantel-Haenszel statistic is presented as a DIF/DPF contrast with DIF/DPF statistical significance, which can be interpreted as having a negligible (< 0.42), moderate ($0.64 > |\mathrm{DIF/DPF}| \geq 0.42$), or large (≥ 0.64) effect on the outcome of the student or item results.[319]

The Rasch-Welch *t*-test is performed by re-estimating the measures of student ability or item difficulty while anchoring the values of the other to the original estimates. For example, if a DIF test is being performed, all the students would be separated into their respective subgroup, and their ability measures would be anchored (locked into place). The item difficulty measures would then be estimated for each student subgroup based on the results of the students and their ability measures. The new difficulty measures obtained from the different subgroups for an item are then compared to each other via a Welch *t*-test (Equation 10, Section 1.6.3) to determine if there is a significant difference between the two measures.

In theory, both the Mantel-Haenszel statistic and the Rasch-Welch *t*-test should produce the same result; however, using the Rasch-Welch is a direct measure of differences whereas Mantel-Haenszel is indirectly measuring the differences. This means that the Rasch-Welch method of measuring DIF and DPF effects is more accurate.[319-324] It is important to note that as the sample size is broken down into smaller subgroups, the effective sample size of each group can be dramatically reduced. The sample size of the subgroup still needs to be large enough to give valid and reproducible results, as

the smaller the sample size the larger the chance of random error that can influence the results of the analysis.

It is possible that DIF can induce the appearance of DIF in other items as an artefact of the DIF identification process.[325] The induced DIF is known as artificial DIF, whereas the DIF that is inherent to an item is known as real DIF. The rate of artificial DIF is dependent of the size of the real DIF, the amount of real DIF, the sample size, and the quality of the assessment. This means that if artificial DIF is found within an assessment it is an indicator that there are potentially items within the assessment that have a large real DIF. Thus, analysing assessments for DIF needs to be an iterative process, whereby the item with the largest DIF needs to either be improved or removed from the analysis before working on the other items that display DIF.

The impact of DIF and DPF on student and item measurement is generally fairly small over the entire assessment, unless the values are large and consistently in one direction (i.e. consistently favour the same subgroup).[309,326,327] However, this does not mean that DIF or DPF does not impact the results of the students, and hence it is still a threat to assessment validity. The use of DIF and DPF analysis is important to ensure that assessments can be considered to be fair to all the students and not show any favouritism to a specific subgroup. However, it is important to consider the context of the assessment, as some assessments expect to see DIF due to the nature of the assessment. For example, an English proficiency test would favour students who grew up speaking English over students for whom English is their second language.

2.6.2 Distractor Analysis

If the answer selections made by the students are included within the raw data inputted into *Winsteps* not only will it generate statistics for the item, but it will also conduct an analysis of each of the answer options. This includes basic statistics on each option, such as the number of times that an option was selected and the percentage of the student cohort that that corresponds to. It also provides the average ability of the students that selected each option, the option's correlation to ability, and fit statistics for each option. This information can be used to identify problematic distractors that might be causing issues within the item.

One important aspect of assessments is that all the options within an item are expected to be functional, unless the item is extremely easy, to ensure that the item is adequately testing the students. The functionality of a distractor (incorrect option) is based on two requirements: firstly, distractors need to be plausible alternatives to the correct answer; secondly, a distractor should see a selection rate that is ≥ 5% of the student cohort.[44,126,224] This is slightly dependent on the item being asked, as it may not be plausible to always meet both of these requirements. However, these two requirements should be the benchmark by which all distractors are judged. The plausibility of a distractor should be evaluated before it is even used within an assessment, while the selection rates of each of the distractors should be evaluated after each use in order to diagnose any issues within the item.

The selection rates of distractors can be used in conjunction with the mean ability measure of the students selecting that distractor to gain a better understanding of the type of students that are selecting each option. For example, if the least plausible option is selected by 5 students who all have low ability measures it is likely that all these students were guessing. In contrast to this, if one of the distractors has a higher ability mean than the correct answer then it implies that there is an issue within the item that is causing the higher ability students to select this option over the correct

answer. While it is impossible to determine exactly what might be causing the issues without further analysis, this result can highlight potential areas that might be causing issues for the students. The correlation measure can also be used in a similar manner, as it informs which options are chosen by higher ability students. It should be expected that all the options, except the answer option, will have negative correlations (i.e. they are more likely to be chosen by students of lower ability). However, it is possible that more than one option has a positive correlation with ability, a result that would be seen from using item analysis. This is not always an indication of a problematic distractor; however, it should be considered further to ensure that it is not causing validity issues within the item.

The infit and outfit values are calculated for every option to determine how well each of them fits the Rasch model. It should be noted that the options will likely follow the trend seen within the item analysis (particularly the answer option). However, this provides information as to which of the options might be causing any of the issues seen within the item. This information can be used to support the conclusions made from other statistical measures, such as what distractors are causing issues and what distractors are functional. Alone, they do not provide much insight without considering them in conjunction with the information provided through other statistics.

One important consideration when looking at the results of the distractors is that the sample size for each option is dramatically less than the entire student cohort. The distribution of the students is also often heavily skewed in favour of a few options that seem the most plausible. This means that some of the statistics produced by the options may be based on very small sample sizes, and hence that needs to be considered when analysing the output of the analysis. For example, it is possible that a high achieving student selects an incorrect option that is only selected by a few other students. This distractor would appear to be dysfunctional, as a result of low levels of student selection rates. However, as a high ability student selected that option it would cause the mean ability and correlation of that distractor to be relatively high, and thus the distractor may appear to be causing problems for high ability students based on those results. It is important to recognise that this is the result of an anomaly, and regardless of whether it was selected by chance or deliberately, the results are not significant enough with such a small selection rate to justify any conclusions. It is also important to remember that it is not the percentage of the student cohort that defines if the statistics are valid, it is the actual number of students. A sample size greater than thirty is enough to assess the items, but that sample size will not be able to generate any meaningful results from distractor analysis, except whether the distractors can be considered functional for that particular assessment.

2.6.3 Anchored Analysis

As discussed previously (Section 1.6.11) while the Rasch model does produce independent measures of student ability and item difficulty these measures are not comparable between assessments. This is because the measures are placed on a logit scale where the zero is determined by the average item difficulty, which means that the zero on the logit scale is different between assessments. However, it is possible to link two or more assessments such that they can be considered to lie on the same logit scale, allowing for the measures from different assessments to be compared directly.[273,309] Linking two or more pieces of assessment requires the assessments to have some amount of commonality that can be used to link the assessments by acting as an “anchor” point for the logit scale. The anchor can be either a student who has undertaken each of these assessments, or it could be an item that is common between the assessments. The more anchor points that can be used, the more accurate the new logit scale will be.[328] Once the common students or items have

been identified, their measures from one of the assessments is used during the estimation of the measures in the other assessment(s). For example, if a specific item was found to have a difficulty of 1.0 logits in one assessment, and that same item was used in another assessment and found to have a difficulty measure of 2.0, this means that relative to the other items used in the assessment the item was harder in the second assessment. However, it is possible to anchor the difficulty measure of this item at 1.0 when estimating the measures of the second assessment, and as a result of this, the logit scale will shift to account for this change and place the other items on the same logit scale used by the first assessment. This allows for the measures of item difficulty and student ability to be compared between the two assessments. This could be used to link assessments within a course to compare item difficulty and results throughout the course, or it could be used between different courses to ensure the items are all testing the students at a similar level.

There are some important considerations when attempting to use anchors to compare between assessments, as it may not always be possible or even reasonable to perform an anchored analysis. The first and most important consideration is that the all the measures on the logit scale still need to uphold the Rasch principle of unidimensionality. For example, comparing student ability measures between an assessment undertaken in a history course and an assessment undertaken in mathematics is not a valid, or useful, comparison. This is because it is highly likely that a student's ability within history does not correlate with their ability in mathematics, as they require different skill sets. This means that the assumption that the anchored students should exhibit the same ability level in both courses is flawed. For the assumption of unidimensionality to be upheld, the two assessments should have a logical connection to each other. This could be as straightforward as the assessments being undertaken within the same course, or as broad as assessments within two different fields of science. However, regardless of the thought process, unidimensionality should always be analysed after performing any anchored analysis to ensure that the assumption of unidimensionality is a logical and mathematically sound conclusion. If unidimensionality it not seen within the analysis then it is highly likely that this is a result of the two assessments assessing different underlying abilities, and thus it is not possible to compare the two measures.

Another important consideration is the selection of anchors, as while it might seem reasonable to simply maximise the number of anchors to obtain the best fitting logit scale this can have a negative impact on the results. This is because any students or items that are skewed as a result of the assessment (i.e. exhibit DPF/DIF) can contaminate and shift the scale as a result of their measures not being an accurate reflection of their true ability/difficulty.[328,329] If the scale is contaminated it can cause other students or items to exhibit DPF/DIF when their measures may be a true reflection of their ability/difficulty. This why it is important to select the students or items very deliberately to be used as anchors to avoid any level of contamination.

To maximise the fit of the logit scales to multiple assessments it is not the percentage of anchors that is important, but the number of anchors. The more items/students used the better the fit will be; however, this also increases the chance of contamination. Thus, typically the number of items/students used is 10, although as little as 4 anchors has been shown to have enough power to place two assessments on the same logit scale.[309,330] Selecting which items/students are used as anchors needs to be based on previous knowledge, expert advice, or an anchor selection strategy.[328-330] Often little is known about which items and students are likely to display DIF and DPF effects, and thus basing anchors on previous knowledge or the advice of experts is unlikely to be a plausible strategy in most cases. This means that a selection strategy needs to be employed to identify the items or students least likely to contaminate the anchors. The following example will focus on

identifying items to be used as anchors, but the process is equally as valid when identifying students to be used as anchors. The first step is to give each item a test statistic that represents the potential DIF strength of that item. This can be done either by anchoring all the other items except for the item being analysed to obtain one test statistic per item (type I statistic), or by anchoring every other item individually generating *k*-1 (where k is the number of items) test statistics for each item (type II statistic). These test statistics can then be used to identify any number of anchor items using the items with the smallest potential DIF strength. In the case of the type II statistic, either the mean test statistic can be used, or a p-value can be equated and used to identify the items with the lowest potential DIF strength. Other potential anchor strategies include using all items as anchors, all items but those of interest as anchors, or iteratively adding or removing anchors based on the items that display DIF after the anchored analysis.

If the change in student or item performance over time is being assessed, where the same students and items are used, it is possible to utilise data stacking or data racking in order to observe these changes without having to anchor the results.[287,331,332] Both of these methods utilise analysing both tests from different time periods within the same dataset. This is done by placing the dataset into a spreadsheet where individual students are represented by rows, and the items by columns. If the two time periods are placed next to each other such that the student has the results from both tests within the same row this is known as racking the data. If the data is racked it means that the change in the item difficulty over time is being analysed, and the student ability is assumed to have remained constant throughout that period. Each item will generate two item difficulty measures that can be compared to observe the change over time, whereas the students will only have one ability measure. Placing the two time periods on top of each other such that every student appears twice, and the items only appear once is stacking the data. Data stacking analyses the change in student ability over time, as each student is assigned two ability measures, while the item difficulty is assumed to have remained the same over that time period and as such is given one difficulty measure. Depending on the data being analysed this can be a much quicker and effective way of measuring changes over time than anchoring. However, the assumptions of whether the difficulty and ability are expected to change need to be seriously considered for each dataset being analysed. This is the way in which anchoring was conducted within this research, as the assumption of equal item difficulty can be justified through item comparison, and it removes the potential for issues with anchor selection as no anchors are required to link multiple assessments.

2.7 Question Breakdown and Comparison

2.7.1 Construction Analysis

Fit statistics and statistical analysis can highlight if there is an issue within an item, but they lack the ability to identify how to fix the issue. There is no technique of analysis that will allow for an issue within an item to be accurately identified. As a result of this, the improvement of items is an iterative process, whereby the problematic items are identified, “improved”, reused in assessment, and then re-evaluated. This process is repeated either until the issue is no longer present, or until the item is removed in favour of a new item. This improvement needs to be undertaken by reviewing the item with the new knowledge obtained from the assessment analysis and by considering the item construction guidelines. There is the potential that the item analysis will have identified a specific distractor option that needs to be corrected; however, if that is not the case the construction of the entire item needs to be broken down and any detail that may cause issues for the students needs to be re-constructed in a way that eliminates that concern. If all the components

of an item are constructed with no apparent issues, then it is expected that the item will be able to assess the students on the desired content. However, students are unpredictable in how they approach and interpret items. As a result, issues such as cueing, student misunderstanding or misinterpretation of the item, and assessing factors unrelated to ability can result from even the most well-thought-out item. This is why it is important to always re-evaluate every item after it is used in an assessment, because while it is possible to minimise the potential of issues arising, it is improbable that every potential issue within an item has been evaluated before it is used within an assessment. Continual evaluation and iteration of items is the only way to ensure that an assessment is testing students on the desired content.

2.7.2 Classifying Multiple-Choice Questions

To develop a system that can be used to classify MCQs, the previously mentioned taxonomies will be used in conjunction with ideas developed within this research in order to develop a technique that can be applied to any assessment. In order to ensure that the technique can be applied to any assessment it will be developed using only one assessment, and will then be applied to other assessments used within this research in order to discover any flaws and missing aspects to the technique. Based on these results, the technique will be iterated and improved continuously until it has reached a point where it can be applied to all the assessments used within this research. The concern with this methodology is that while the iterative method helps improve the technique, it only accounts for the types of items that appear within Biology and Chemistry assessments, as these are the assessments used within this research. This means that it is possible that this method of classification will be missing some aspects that appear in items from different topic areas. However, with more iteration using items that do not appear to fit within the technique, it should be possible to construct a method of classification that can be used on all MCQs.

Chapter 3: Assessment Tasks and Items

3.1 Section Outline

3.1.1 Research Questions

The first step in any data analysis project is to ensure that the methods being used are valid and appropriate for the data that is being analysed. Using the methods described in Chapter 2 will allow for two of the research questions to be addressed within this section and allow for statistical testing in future sections:

Are the MCQ items used at The University of Adelaide in first-year Chemistry courses performing as they are expected to?

Is there a significant difference in the performance of male and female students within MCQ assessments? If so, how can this be addressed?

In this section, both of these questions are evaluated using Classical Test Theory (CTT) and Rasch analysis to illustrate two different methodologies that can be employed to reach conclusions on the validity of an assessment and its items. Both methodologies evaluate the functionality of the assessment task and the performance of each item within that assessment, which allows for conclusions to be drawn as to whether they are behaving as expected. In this research the assessment tasks will be analysed first to highlight if there are specific areas of concern before the individual items are analysed. These methodologies can also be used to compare the relative performances of the male and female student cohorts through different analytical approaches.

3.1.2 Project Objectives

Due to the close correlation between the research questions and the project objectives, the analysis described in this section will be used to address several objectives that are closely related to the identified research questions. Those two objectives are:

To assess the items used in MCQ assessments both currently and previously at The University of Adelaide in first year Chemistry courses to determine whether the performance of the students on these items is providing assessors with information that reflects the ability of the students on the content being assessed

To compare the performance of male and female students in first year Chemistry MCQ assessments at The University of Adelaide to ensure that any difference in performance is a result of a difference in ability and not due to factors within individual items that influence student performance based on student gender

The completion of these objectives allows for the determination of whether there is either (or both) a set of items that show deviations from how they are expected to perform within the assessment and a set of items that display differences in the performance of male and female students. Should such individual items be found, then analysis of the construction of the items can be undertaken to determine what can be done to improve them. Such item analysis also provides the opportunity to

attempt to classify those items in order to uncover trends and aspects within item construction that may influence student outcomes in particular ways. Undertaking this classification would allow for the completion of another research objective.

To analyse the construction of MCQ items utilised at The University of Adelaide in first year Chemistry courses to develop a methodology for the classification of MCQ items

Through the completion of this objective it is hoped that a process can be generated that other assessors can apply to their own assessments that would allow the identification of individual items of potential concern and provide confirmation that the construction of the assessment matches the purpose of the assessment.

3.2 Analysis of Assessment Tasks

3.2.1 Exploratory Analysis

The first step of any assessment task analysis is determining the how the cohort of students performed in the assessment itself, which can be done by considering basic descriptive statistics of the average results and the spread of those results. This includes the number of students who undertook the assessment, the mean result for the cohort of those students, the standard error of those results, and the standard deviation of the cohort. This initial analysis was performed for every assessment task that was considered in this research; an example of this analysis is tabularised as shown below in Table 3 (see Appendix 7.1 for the statistical results of the other assessment tasks analysed).

Table 3: MCQ Assessment Results from Chemistry IA 2012

	n	**Mean Score**	**Std. Deviation**	**Std. Error**	**Max. Score**	**Mean %**
Lecture Test 1	471	8.76	3.214	0.148	15	58.4
Lecture Test 2	446	8.50	2.646	0.125	15	50.0
Exam Part 1	508	4.61	2.099	0.093	10	46.1
Redeemable Exam	488	16.09	7.395	0.335	30	53.6

There is some information that can be easily observed from this table, such as the average result was often slightly above 50% for most of these assessments and that more students undertake the Exam Part 1 assessment task than the other assessments (likely due to the redeemable nature of the other assessments). It can also be observed that the standard deviation and the standard error is higher in the assessment tasks that contained more items, which is unsurprising as this allows for a larger range of potential student results. These trends are consistent across all of the assessment tasks being analysed; however, there are several questions that cannot be answered from these results. For example, this does not show how distinct groups within the cohort are performing; it does not inform the distribution of the student results; and it provides minimal insight into how students are approaching the redeemable nature of the assessments. While it is not possible to answer some of these questions without a deeper analysis of the data, the distribution of the student results can be observed using histograms and generating measures of the spread. It is expected that the results of the students follow a Normal distribution based on the concept of the Central Limit Theorem, as each student provides their own independent measure to a much larger sample size. Figure 4 displays a histogram of the results of students in Lecture Test 1 from Chemistry IA (2012) (see Appendix 7.2 for the histograms of all the assessment tasks analysed). The figure

illustrates that while the data does not follow a perfect Normal distribution there is a high level of correlation between the results and a Normal curve. One of the issues that needs to be considered when comparing assessment results to a Normal distribution is that ideally the distribution has a large range such that the sample will naturally tail off at each end. This is not completely realistic in assessments as the spread of the data is restricted to the number of items asked of the students, which is restricted by the time frame placed on the assessment. The fewer items used within an assessment the greater this problem becomes, as less information is being obtained about the students and their ability, which may impact the effectiveness of the assessment fulfilling its purpose. Therefore, looking at the distribution of the students can be used to both judge how well the assessment is able to separate the students based on their performance, as well as inform initial impressions about the performance of the items and their ability to identify a range of student abilities. If the distribution of the results was heavily skewed in one direction it would imply that either the items being used were all too easy or all too hard for the students, and thus no effective information about the relative performance of the students would be obtained from that assessment. Thus, except on MCQ assessments with a large number of items, it should be expected that the distribution of the students' results is likely to be slightly shifted to the right because more students will be achieving higher marks than the Normal model expects due to the nature of assessment construction.

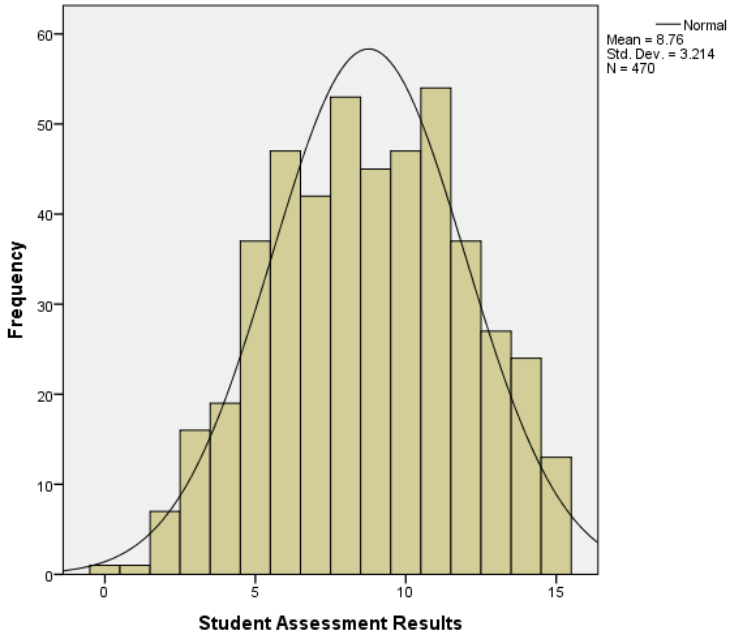

Figure 4: Student Scores Obtained in Chemistry IA Lecture Test 1 2012

The measures of spread can be used to describe the shape of the graph and for this the measures of skewness and the kurtosis of the distribution were considered to ensure that they were within reasonable ranges for what is expected of a Normal distribution. An example of the measures of the spread of the data can be seen below in Table 4 (see Appendix 7.1 for the measures obtained for all other assessment tasks analysed).

Table 4: Measures of Spread of MCQ Assessments Undertaken in Chemistry IA during 2012

	Skewness		Kurtosis	
	Value	**Std. Error**	**Value**	**Std. Error**
Lecture Test 1	-0.067	0.113	-0.709	0.225
Lecture Test 2	-0.290	0.116	-0.569	0.231
Exam Part 1	0.211	0.108	-0.364	0.216
Redeemable Exam	-0.420	0.111	-0.583	0.221

The skewness of the distribution is a measure of symmetry within the data, and the kurtosis is a measure of whether the data has a large or small tail in contrast to a Normal distribution. Ideally the skewness measure lies between -0.50 and 0.50 as this indicates that the distribution is symmetrical as required for a Normal distribution.[237] The kurtosis measure is required to lie between -3.00 and 3.00 for the data to have tails that are reasonable for a Normal distribution.[237] Analysing the results seen within Table 4 it can be observed that the results of all of these assessment tasks lie within the ranges stated, which implies that they tend to follow a Normal distribution.

While looking at the histogram for each assessment utilised shows a strong resemblance between the data and a Normal distribution it is possible to test the data for normality. This was done using the Kolmorov-Smirnov test and the Shapiro-Wilk test to determine if the data deviates significantly from a Normal distribution. An example of this analysis can be seen below in Table 5 (see Appendix 7.1 for the normality tests for all other assessments analysed).

Table 5: Tests of normality of MCQ Assessments Undertaken in Chemistry IA during 2012. Highlighted Cells indicate Observation of a Statistically Significant Difference

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	**df**	***p*-value**	**Statistic**	**df**	***p*-value**
Lecture Test 1	0.086	470	<<0.001	0.979	470	<<0.001
Lecture Test 2	0.104	446	<<0.001	0.971	446	<<0.001
Exam Part 1	0.114	508	<<0.001	0.973	508	<<0.001
Redeemable Exam	0.079	488	<<0.001	0.954	488	<<0.001

Looking at Table 5 it can be seen that both tests of normality found the distribution of the data to be statistically significantly different from a Normal distribution (p-value < 0.05); however, whenever doing these sorts of tests it is important to consider the sample size being used. At larger sample sizes these tests lose their functionality as a method to test for normality because with a high sample size even a small variation from the expected Normal distribution will appear statistically significant within these tests.[269] Therefore, instead of relying on the normality tests, the skewness and the kurtosis of the data was used to determine if the distribution of the data functionally followed a Normal distribution.

Both of these measures (skewness and kurtosis) are also influenced by the sample size; however, the expectations of the measures can be changed based on the size of the sample. The larger the sample size, the greater the allowed values of skewness and kurtosis before the distribution is thought to no longer represent a Normal distribution.[333] The issues with using mathematical methods to determine normality is that it is hard to visualise what that means in terms of how the data is deviating from a Normal distribution and where the difference in expected values and observed values occurs. To generate that visual, a Q-Q plot was constructed, which can be used to show where the data is

deviating from the expectations of the Normal model. In Figure 5 an example Q-Q plot demonstrates that most of the deviations lie at the tails of the dataset (the Q-Q plots for the other assessment tasks analysed can be seen in Appendix 7.2).

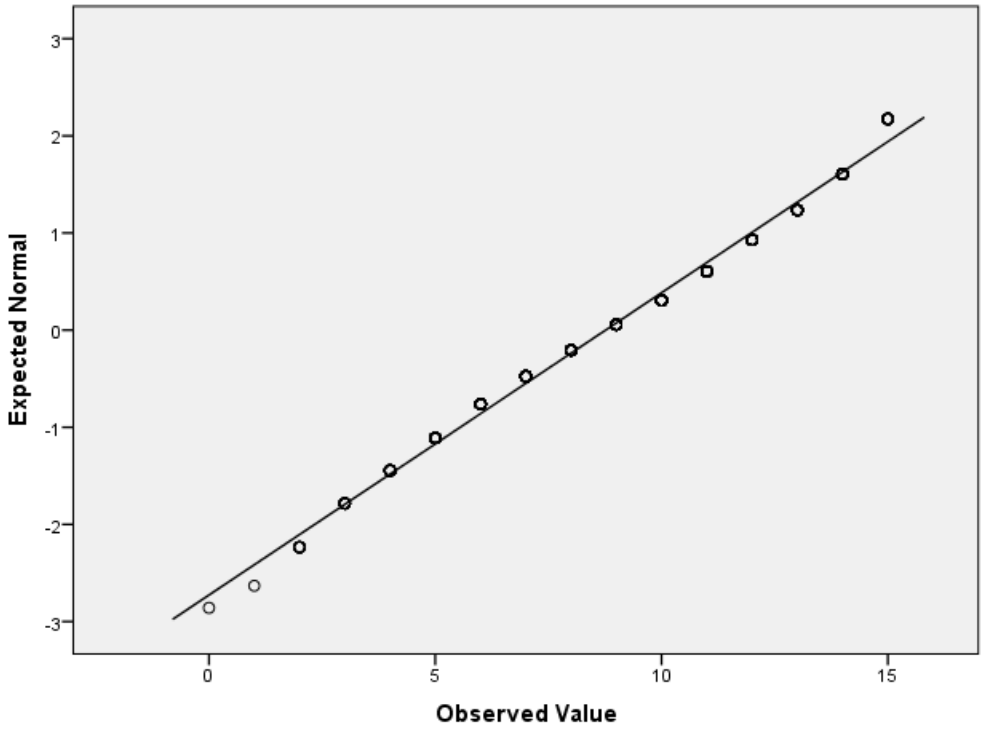

Figure 5: Q-Q Plot of the Student Results in Chemistry IA Lecture Test 1 in 2012

It is not unexpected that there is some deviation from the Normal model at the higher and lower scores within an assessment. This is because it is less likely that students will obtain a minimum score in an assessment (particularly an MCQ assessment) than is predicted by a Normal distribution of the data, as the assessment is likely to include some easier items that the students will be expected to answer correctly. Similarly, it is more likely for students to obtain the maximum score in an assessment than what is predicted by a Normal distribution if the number of items being asked of the students is not large enough to assess the full breadth of their knowledge, which is not always feasible in a timed assessment setting. Analysing the results of the tests of normality across all of the assessment tasks shows that they all fail both statistical tests of normality, except for one assessment task which only fails one of the two tests. However, of the 64 assessment tasks considered in this study all but eight assessment tasks show measures of spread and distribution (skewness and kurtosis) that fit within ranges that are expected of distributions that match a normal distribution. The eight assessment tasks that did not fit within those ranges only showed very minor deviations within the skewness of the data (skewness between -1 and -0.5, or between 0.5 and 1), but not the kurtosis. In addition to reviewing the Q-Q plot of each assessment task allowed for the assumption of normality within the dataset to be justifiably satisfied for every assessment task being analysed within this research. While this in and of itself does not provide a large amount of knowledge about the dataset, it does justify some of the analysis that will be performed on the dataset at later stages in the research.

3.2.2 Classical Test Theory

When evaluating an assessment task the two factors that classical test theory (CTT) allows for the calculation of are whether the student performance is correlated across the entire assessment, and how well the results of the students are distributed across the assessment. The results of these calculations for a full semester of Chemistry IA can be seen in Table 6 (for the results for all the assessments being analysed see Appendix 7.3). The closer these values are to 1 the more ideal the assessment is.

Table 6: Results of a CTT Assessment Level Evaluation for MCQs Undertaken in Chemistry IA 2012

	KR-20	**Ferguson's Delta**
Lecture Test 1	0.7361	0.9758
Lecture Test 2	0.5809	0.9545
Exam Part 1	0.5383	0.9535
Redeemable Exam	0.8276	0.9819

The results of the Ferguson's Delta test clearly show that all the assessment tasks had a high level of student distribution ($\delta \geq 0.90$), which could be expected based on the distribution seen in the histograms. There is some amount of variance within the correlation of the student performance across the assessment as measured by the KR-20 test (ideally $r_{test} \geq 0.70$), which could be caused by several factors. Commonly, assessments such as the ones being analysed here cover more than one course topic per assessment, and thus if the topics are different enough it is possible that the students may perform differently in one topic than another, resulting in a poor correlation of student performance across the assessment. Another consideration is the number of items being used in the assessment, as one of the issues with CTT as a method of analysis is that it is dependent upon the size of the student cohort and the number of items being asked within the assessment. The larger the student cohort, the higher the expectation is that there will be a high level of distribution amongst the results of the students. For the assessment tasks under consideration here the student cohort typically lies between 300 – 500 students depending upon the course, and thus in the same way that the distribution will approximate a Normal model at a high enough sample size it should be expected that Ferguson's Delta will be high if there is a large number of students who sat the assessment. The number of items present within the assessment will influence the amount of correlation seen across the assessment, as a single item that misfits with the rest of the assessment will have a larger influence if less items are used. Conversely, if there are two highly distinct topics covered within the assessment, having more items will make that difference more apparent; hence, it will be easier to distinguish between the results of the topics. The idea that the number of items is important to the statistics can easily be seen within the results, as the redeemable exam assessment is constructed with 30 items, each lecture test contains 15 items, and the exam Part 1 only has 10 items. Knowing that makes it clearer as to why the redeemable exam has the highest correlation across the assessment, while the Exam Part 1 shows the lowest correlation. This trend continues across all the assessment tasks analysed, where typically the Exam Part 1 in every year has the lowest correlation while the redeemable section within the exam has the highest. It is important to remember such considerations when approaching these types of analyses, as just because one assessment has a higher correlation it does not necessarily make it the best assessment. For example, it is quite common that the redeemable section within the exam is constructed by reusing a large proportion of the items that were asked within lecture test 1 and 2, but the metrics produced for the redeemable assessment look better despite the fact that the same items were used simply because more items were in the assessment. What this means is that the metrics produced here

should be evaluated, but the circumstances around them need to be considered, as the reasons for the shifts in the values is not always reflective of the assessment itself. Despite this, based on all of the assessment tasks analysed there were no specific problems highlighted for any one particular assessment task; however issues observed within an entire assessment task tend to only be seen if they are completely undermining the purpose of the assessment task causing the assessment to deviate from its original purpose.

3.2.3 Rasch Analysis

The second methodology used in this research to evaluate assessment tasks was the Rasch model, which analyses both the students and the items based on how reliable their measures are and how well the measures separate. The reliability of the measures simply relates to how reproducible they are: the more reproducible they are the more accurate the measures must be. The separation of the measures determines how well the assessment can establish a hierarchy within the measures to distinguish between different tiers of performance. An example of the measures produced by a full semester of assessment can be seen in Table 7 (see Appendix 7.3 for the results from all of the assessments analysed) which shows how effective the assessment is at determining student ability and item difficulty measures.

Table 7: Rasch Measures for Assessment Level Analysis of MCQ Assessments in Chemistry IA 2012

	Item Reliability	**Item Separation**	**Student Reliability**	**Student Separation**
Lecture Test 1	0.98	7.19	0.68	1.47
Lecture Test 2	0.99	8.25	0.55	1.10
Exam Part 1	0.98	7.46	0.50	1.01
Redeemable Exam	0.99	8.43	0.82	2.12

Based on these measures, the item separation and reliability for all the assessments are well above what is expected to achieve reasonable measures (item separation > 3, item reliability > 0.90).[273,309] This implies that all of the item measures are reproducible and it is easy to distinguish between items of varying difficulties, which should be expected when the number of students undertaking an assessment is high as it provides a large amount of information about the relative difficulty of the items. This trend within the item measures for the assessment tasks is true across all of the assessment tasks being analysed within this research, which means that there is a high degree of confidence in the item measures throughout this analysis. In contrast to this, the student reliability and separation measures are much lower (expected values: student separation > 2, student reliability > 0.80).[273,309] This implies that the assessment tasks may be quite poor at being able to separate the different levels of student performance, and the results may not be reproducible if the same students sat the assessment again. The student reliability and separation were also highly consistent throughout all of the assessment tasks analysed, where they are usually slightly below what is expected of them. The reason for this is the same reason that has been highlighted in the previous assessment level analysis, which is that there are not enough items present within the assessments to provide enough information about the performance of the students to be confident that the students will always fall into the same hierarchy. This can be observed by comparing the separation and reliability calculated for the redeemable exam to the other assessments being analysed, as the key notable difference between those assessments is the number of items and not strictly the actual items themselves. This suggests that the actual construction of the assessment itself is not an issue, but rather the amount of information it is providing about the students due to the number of items present within the assessment is lacking, and thus impacts the confidence that

can be placed within the results of the assessment. However, an important consideration is that the student's final grade is never decided on a single piece of assessment alone, and thus it is reasonable to suggest that even though the results produced by each individual piece of assessment may not be completely reflective of the student's ability, using the combined results of all of the assessments will form a more complete picture of where the ability of each student lies.

Another consideration that can be measured through the Rasch model is whether the assessment is unidimensional, or whether it is multidimensional and requiring the students to be competent with multiple distinct skill sets. This is important because the Rasch model assumes that the ability and difficulty measures alone give enough information to predict the performance of the students. If there are other factors present within the assessment that are not accounted for within the model then this assumption is obviously flawed, and thus the Rasch model cannot be used to explain the performance of the students and the items. Similarly, if the assessment is not unidimensional, it could imply that any issues within the assessment are not due to the content of the items, but rather other underlying factors within the assessment. Rasch modelling tests unidimensionality by using the residual variance to determine if any additional factors have enough influence on the results to be considered as another dimension within the model. An example of this can be seen in Table 8 (see Appendix 7.3 for the results from the other assessments analysed). Typically, for a measure to be considered a statistically significant influence on the performance of the students or items it needs to have an eigenvalue greater than two, otherwise it is considered to be within the noise level of the data.[273,309]

Table 8: Dimensionality Test for the MCQ Assessments used in Chemistry IA 2012

	Measures Eigenvalue	**Measures %**	**1st Contrast Eigenvalue**	**1st Contrast %**
Lecture Test 1	5.5164	26.9	1.3521	6.6
Lecture Test 2	5.4719	26.7	1.7480	8.5
Exam Part 1	3.4318	25.5	1.5241	15.2
Redeemable Exam	11.5157	27.7	1.7902	4.3

It can be seen by looking at Table 8 that the eigenvalues for the ability and difficulty measures (considered together as the measures eigenvalue) are well above the noise level (> 2), and thus statistically they contribute significantly to accounting for the residual variance within the data. Looking at the 1st Contrast (represents the largest factor that contributes to the residual variance) all the eigenvalues fall within the noise range (<2), and thus it is reasonable to state that, for the purpose of this analysis, the assessment is unidimensional. Only the 1st Contrast is shown within Table 8, even though multiple other contrasts are measured, as each successive contrast shows less and less significance in its contribution to the residual variance. Thus, if the first contrast lies within the noise level then every successive contrast will also lie within the noise level, as the 2nd Contrast and every successive contrast will never be larger than the first. The dimensionality tests from all of the assessment tasks being analysed showed that the assumption of unidimensionality was valid across all the assessment tasks, and thus all the assessment tasks are only assessing the students in one ability. It is assumed that the ability being assessed is the student's ability within chemistry and the topics being covered by the assessment; however, that can be neither confirmed nor disproved by these results. It is highly unlikely that the assessments could be heavily influenced by some other ability and not at least have a single contrast above the noise level that represents the chemistry ability of the students. This means that it is a justified assumption that chemistry ability is the

unidimensional trait that is being assessed, but it is important to remember that this is an assumption that can never be confirmed.

A Wright map can also be created using Rasch analysis, which generates a visual representation of the entire assessment and the comparative measures of the students and items within it. The Wright map does include the individual student and item measures within it; however, more importantly it shows how the ability measures of the students compare to the difficulty of the items. This allows for the assessment to be evaluated based on how well it is targeting the ability of the students, as when the student ability and item difficulty are equal the students are predicted to have a 50% probability of correctly answering that item. This means that the most information is obtained about the students when the item difficulty closely matches the student's ability, as whether the student lies above or below a level can be determined using those items. Therefore, in an ideal assessment the difficulty of the items should be normally distributed across all the ability measures of the students to obtain the most information about the students based on their performance in the assessment. Looking at a Wright map from one of the assessments analysed Figure 6 (see Appendix 7.4 for the Wright maps of all of the assessments analysed) shows that the distribution of student ability measures does seem to follow a roughly Normal distribution in the same way that the raw scores of the students did. It can also be seen that while the item difficulties do not strictly follow a Normal distribution (mostly due to the number of items present within the assessment) they are spread across a wide range, which ensures that the performance of the students in the assessment gives information about their ability level regardless of where they lie on the scale. It is only at the extreme ends of the scale that no information about the students is being learnt; in Figure 6 this refers to the cluster of students that lie two standard deviations above and below the mean. It is expected that these students obtained either full marks or no marks within the assessment and thus the model cannot accurately place them in comparison to the other students. It is important to remember that the purpose of assessment does not require the ability level of every student to be known, as commonly it is used to determine the progress of the students in their learning. Students who obtain full marks on assessment clearly demonstrate that they are above the standard that is expected of them, and thus it is not important exactly where their ability lies in relation to the rest of the student cohort. Similarly, students who obtain no marks show that they are not progressing in their learning at a rate that is expected of them. Thus, Wright maps were used to ensure that the assessment is constructed in such that way that the items are obtaining relevant information about the ability level of the students.

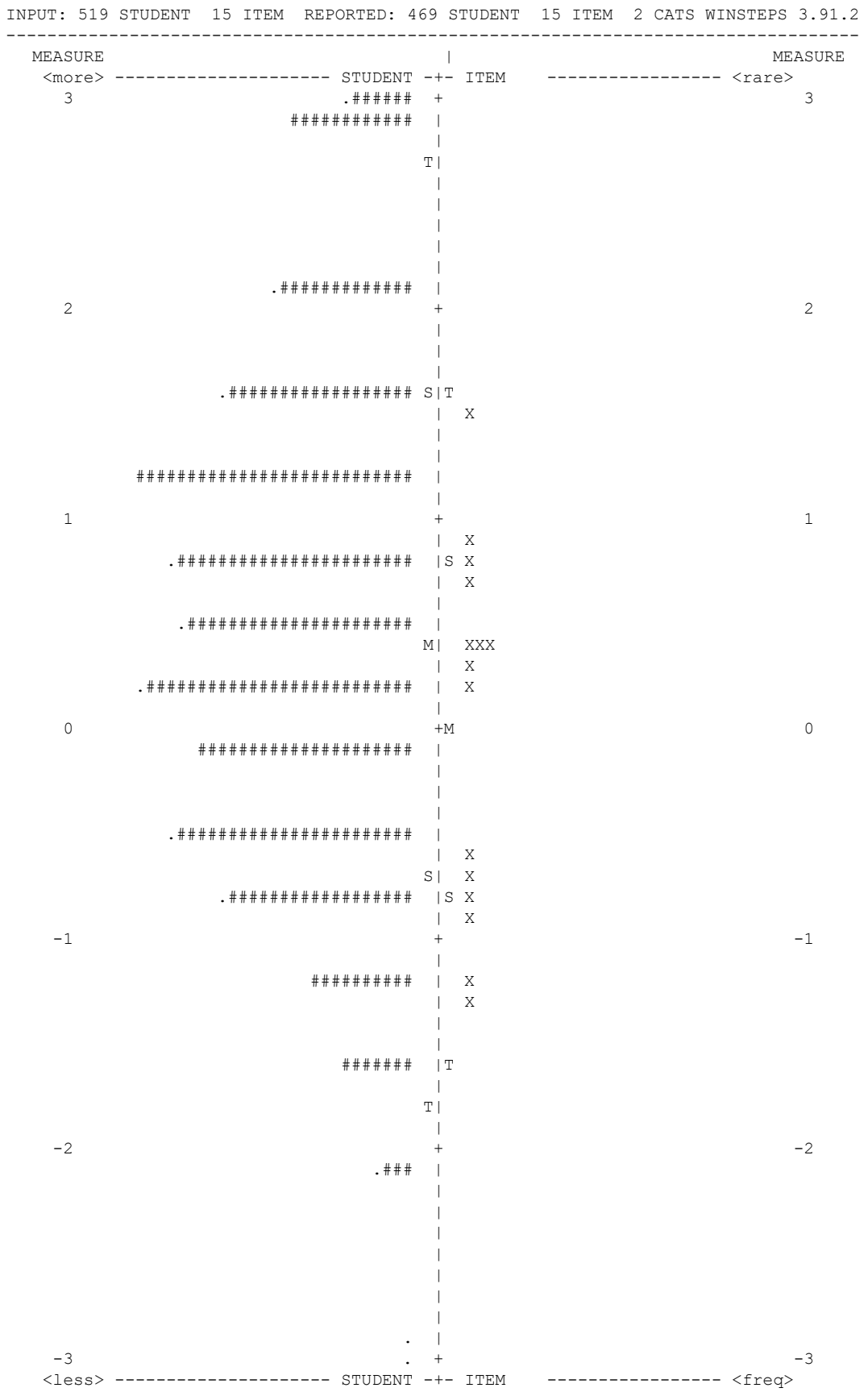

Figure 6: Wright Map of Student Ability and Item Difficulty from Lecture Test 1 in Chemistry IA 2012

3.3 Determining the Performance of Individual Items

3.3.1 Classical Test Theory

Once the assessment tasks have been analysed it is important to break down how individual items are performing within those assessment tasks, as any items that are not performing as expected will impact upon the validity of the entire assessment. This is particularly important if an issue has been identified within the overall analysis of the assessment task, as the only way to fix those concerns is to identify which items are the root cause of the problem; however, it is possible that there are flaws within the items that do not cause issues within the analysis of the assessment tasks. The simplest methodology used to analyse the items was Classical Test Theory (CTT), which gives measures of item difficulty, item discrimination and item correlation between the results obtained on the individual items to the results of the assessment task as a whole. These three measures were generated for every item, and then each of them was evaluated based on the values that they are recommended to lie between. An example of this evaluation can be seen in Table 9 (see Appendix 7.5 for the values obtained for each item used in the assessment tasks analysed), where each aspect needs to be considered both individually as well as how it relates to the other measures obtained through the analysis.

Table 9: Classical Test Theory Analysis of the Items used in Lecture Test 1 in Chemistry IA 2012

Item	Item Difficulty (P)	Discrimination Index (D)	Correlation (r_{pbi})
Lec_1_1	0.749	0.340	0.450
Lec_1_2	0.808	0.196	0.403
Lec_1_3	0.421	0.553	0.508
Lec_1_4	0.540	0.383	0.390
Lec_1_5	0.796	0.153	0.328
Lec_1_6	0.696	0.366	0.438
Lec_1_7	0.511	0.553	0.522
Lec_1_8	0.715	0.443	0.514
Lec_1_9	0.506	0.553	0.542
Lec_1_10	0.323	0.426	0.432
Lec_1_11	0.436	0.375	0.416
Lec_1_12	0.730	0.366	0.497
Lec_1_13	0.555	0.409	0.425
Lec_1_14	0.460	0.536	0.505
Lec_1_15	0.521	0.477	0.443

To obtain the most relevant information it is important that the difficulty of the item is neither too hard nor too easy for the students, that the item is clearly able to differentiate between students of varying ability level, and that a student's performance on an item matches their performance on the assessment. Analysing Table 9 shows that all the items in this piece of assessment have reasonable values for both their difficulty ($0.30 < P < 0.90$) and correlation ($r_{pbi} > 0.20$); however, item Lec_1_2 and Lec_1_5's discrimination index score is lower than what is expected of an ideal item ($D > 0.30$). Another way of viewing the items is by graphing the different CTT item values against each other to observe the measures of the entire assessment tasks together so that outliers can be easily noted. This can be seen within Figure 7 and Figure 8 below, which show item discrimination and item correlation plotted against item difficulty to highlight items performing abnormally.

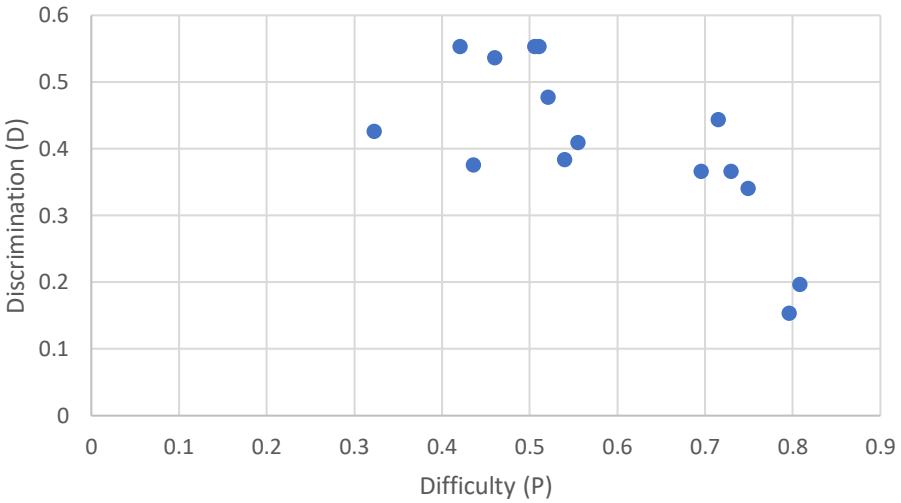

Figure 7: Classical Test Theory Item Difficulty versus Item Discrimination for Lecture Test 1 from Chemistry IA in 2012 to Identify Problematic Items and Evaluate the Assessment Task

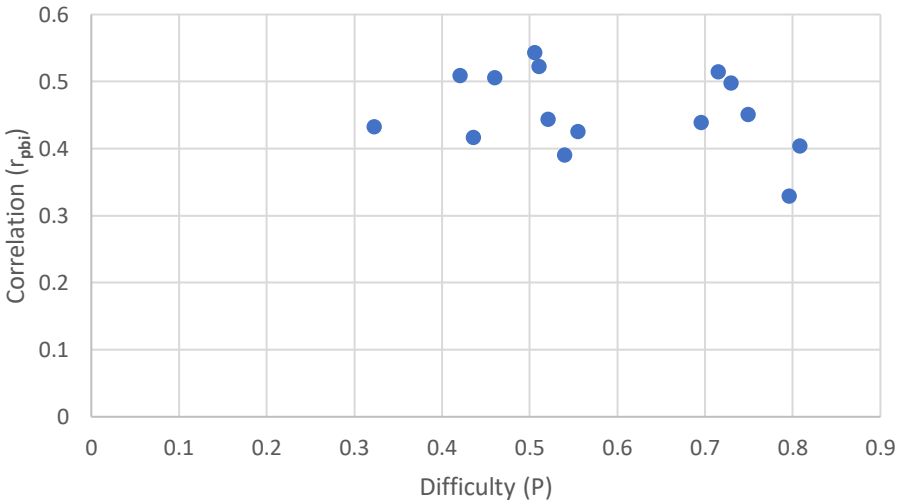

Figure 8: Classical Test Theory Item Difficulty versus Item Correlation for Lecture Test 1 from Chemistry IA in 2012 to Identify Problematic Items and Evaluate the Assessment Task

Reviewing the two figures clearly shows the two items that were previously identified to have a discrimination value below what is expected of a normal functioning item. Other than those two items none of the other items show any issues, and thus by reviewing these figures for each assessment task it is easy to identify if the task is suffering because of the items. In this instance even though those two items have low item discrimination values the fact that there are no other issues present within the assessment task means that the entire assessment task is expected to be able to fulfil its purpose. It can be reasoned that in both cases the item discrimination levels are lower due to the difficulty of the items. The higher the item difficulty, the more students obtained the correct answer on that item (i.e. an item difficulty of 0.808 means that 80.8% of the student cohort obtained the correct answer on that item). Thus in the case of both of the items with low discriminatory values, a large percentage of the student cohort was able to correctly answer the question, and hence it should be expected that the item is not able to clearly show a difference in the performance of high ability students compared to low ability students when both provide the

correct answer. This highlights the importance of considering all the values provided by the analysis and how they inform the expectations of the item performance before judging whether an item is not performing as it is expected to. An example of a poorly performing item can be seen in Table 10, where all three of the item measures characterise the item to be poorly performing.

Table 10: Problematic Classical Test Theory Item from Lecture Test 2 in Chemistry IA 2012

Item	Item Difficulty (P)	Discrimination Index (D)	Correlation (r_{pbi})
Lec_2_9	0.150	0.045	0.008

It can be observed for this item that the students find it difficult (P < 0.30), it does not discriminate between high performing and low performing students (D < 0.30), and it does not correlate with the final results of the assessment (r_{pbi} < 0.20). This indicates that this particular item is found to be difficult by the students for reasons that are likely to be completely unrelated to the topic being assessed, and that there may be an element of guessing involved as lower ability students are just as likely to provide the correct answer as higher ability students within this item. Thus, this item needs to be either reviewed or removed from the assessment to ensure that it is not adversely affecting the results of the students for reasons that are completely unrelated to the purpose of the assessment.

By carrying out CTT analysis on four years' worth of data across the four different first year chemistry courses offered at The University of Adelaide it was determined that out of the 261 unique assessment items used, 12 of them had consistent problematic measures that occurred on more than one occasion (The 12 items and the values associated with each problematic occurrence of the item can be seen within Appendix 7.8). Based on the CTT measures produced for each one of these items it can be theorised that four of these items only contain minor issues (small, but consistent issues within one or two of the CTT values), four of them are likely to contain major issues (large and consistent issues observed within two or three of the CTT values), and four of them require further analysis to determine if they are minor or major issues (usually these items appear problematic due their low difficulty value [i.e. meaning they are hard items for the students] and it is unclear if their difficulty is due to the content or item construction without further analysis). This means that based on the results of CTT there are only 4 unique items out of 261 that may pose a threat to assessment validity; however, one of these alone is not enough to invalidate an entire assessment task. Therefore, based on the analysis of CTT the assessment tasks are expected to be providing accurate information about the students with only 4 items that need serious consideration, and 8 items that may require minor tweaks or clarifications made to their construction.

3.3.2 Exploring Rasch Analysis Measures

The other way that individual items were analysed within this work was through the use of Rasch modelling, wherein it is expected that each assessment item fits the expectations of the model otherwise it is classified as misfitting and potentially needs to be evaluated or removed. This is because misfitting items do not meet the expectations of the Rasch model, and thus as Rasch analysis is a confirmatory model it means that items that do not meet its expectations are seen as items that are the most likely to be causing any issues within an assessment task. This analysis was undertaken on all the MCQ assessment tasks utilised in first year chemistry courses at The University of Adelaide over the four-year period of 2012 – 2015. As applying the Rasch model generates new measures from the data of student ability and item difficulty as part of the analysis those measures

also needed to be evaluated in the same way as the raw scores to determine the average values and their spread. The item sample size is small in all of the assessments (Lecture Test: 15 items, Exam Part 1: 10 items, Redeemable Exam: 30 items), except when the assessments are combined together, which means that there is a large amount of error within both the tests of normality and the descriptive measures of the distribution. If the assessments are combined to report on the results of all the MCQ assessment tasks carried throughout the semester there may be a concern that the measures may be influenced by the assessments taking place at different times. However, as Rasch measures of student ability and item difficulty are independent from each other, the difference in student performance due to timing differences will not affect the item difficulty, which is not expected to change over time. Combining the results of a student across the entire semester is only a concern as the MCQ assessments are redeemable, and thus their ability measure may not be reflective of the result they obtained at the end of the semester. Despite this, it should be expected that the ability measure generated for the students is reflective of their ability within the course, and thus can still be considered to be a valid way of analysing the students. Even though this methodology can be justified, the potential influence that combining the assessment tasks may have on the measures still needs to be remembered throughout the analysis. An example of the distribution of the item difficulty measures produced for an assessment task can be seen below in Figure 9.

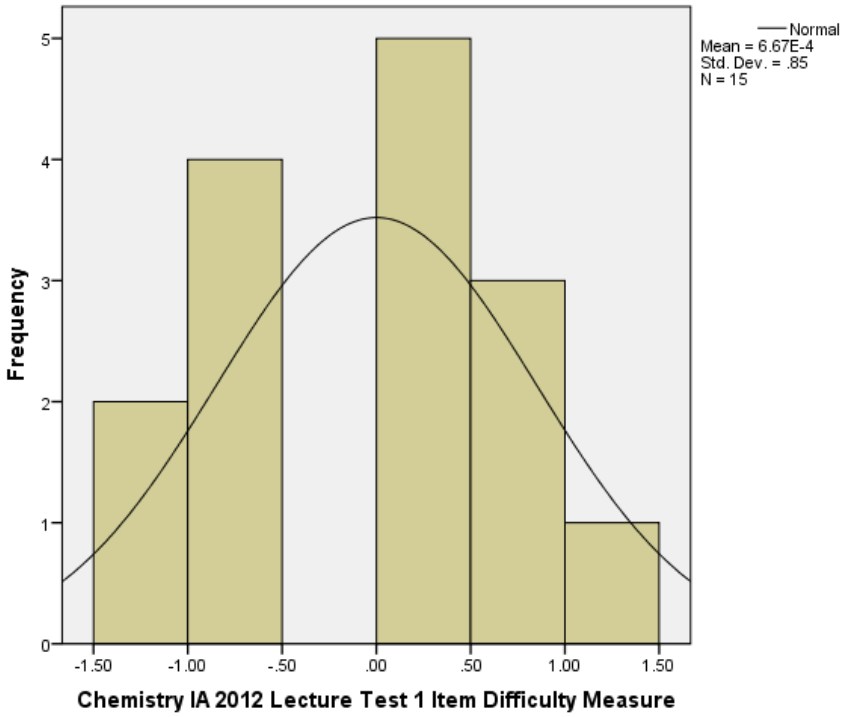

Figure 9: Histogram of the Rasch Item Difficulty Measures in Lecture Test 1 from Chemistry IA 2012 to Determine the Distribution that the Measures Follow

The sample size (*n*=15) makes it hard to distinguish if the data follows a Normal distribution closely enough that it approximates it, which is why the measures of spread are also analysed. This can be seen by the values given in Table 11 (see Appendix 7.6 for the distribution measures of all of the assessment tasks).

Table 11: Measures of Spread of Rasch Item Difficulty Measures in Chemistry IA 2012 Assessments

	Skewness		Kurtosis	
	Value	**Std. Error**	**Value**	**Std. Error**
Lecture Test 1	-0.122	0.580	-1.153	1.121
Lecture Test 2	1.344	0.580	1.294	1.121
Exam Test	0.249	0.687	-0.605	1.334
Redeemable Exam	1.009	0.427	1.906	0.833
Combined	0.715	0.287	0.585	0.566

Using the values of skewness and kurtosis seen within the table it can be seen that while there is no concern for how the data trails off on each side (as measured by the kurtosis) there is some concern over the symmetry of the distribution within some of the assessment tasks. The positive values indicate that the data tends to skew to the right, which implies that there are more items with higher difficulty measures within the assessment task. The larger skewness values (values > 0.50) suggest that the data may not follow an approximately Normal distribution, which may undermine some of the statistical test that assume a Normal distribution, and thus the tests of normality need to be analysed to ensure that the assumption of normality is justified. An example of the application of such tests can be seen in Table 12 (see Appendix 7.6 for the tests of normality on all the assessment task item difficulty measures).

Table 12: Tests of normality of Rasch Item Difficulty Measures in Chemistry IA 2012 Assessments. Highlighted Cells indicate Observation of a Statistically Significant Difference

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	**df**	***p*-value**	**Statistic**	**df**	***p*-value**
Lecture Test 1	0.184	15	0.186	0.936	15	0.333
Lecture Test 2	0.312	15	<<0.001	0.839	15	0.012
Exam Test	0.180	10	0.200	0.928	10	0.427
Redeemable Exam	0.147	30	0.095	0.941	30	0.099
Combined	0.091	70	0.200	0.963	70	0.035

The tests of normality show that all the assessment tasks except for Lecture Test 2 are considered to match a Normal distribution, which means that the assumption of normality is justified for the purposes of conducting statistical tests. Lecture Test 2 appears to be too skewed to the right to match a Normal distribution, and needs to be considered when statistical tests are being applied; however, usually the stacked datasets are used when conducting analysis on the item measures. Across all of the assessment tasks under consideration in this study (64) there are only six tasks that fail the tests of normality (two of which are shown above) and only two (Lecture Test 2 in Chemistry IA in 2012 and 2013) of them fail both tests (these also both show a major deviation from the expected skewness (-1 < skewness > 1)). Those two assessment tasks are the largest concern and need to be monitored throughout the rest of the analysis to ensure that they do not cause issues for the statistical tests; however, it is expected that due to the way in which comparisons are made that this will not cause an issue as they are not reliant on the items from those assessment tasks. There are three other assessment tasks that have major skewness deviations, and one that has a kurtosis deviation (-2 < kurtosis > 2), but on those occasions it is likely that the smaller sample size is causing issues for generating the measures of spread as all of those assessment tasks pass the tests of normality. There are also twenty-one minor skewness deviations (skewness between -1 and -0.5 or

between 0.5 and 1), but once again these assessment tasks showed no issues within the tests of normality. Therefore, while the results that some of the item difficulty distribution of specific assessment tasks is a concern for applying statistical tests, as long as this result is considered, it is unlikely that this will cause any issues within the statistical tests, particularly as all the other assessment tasks show no large issues.

While the item sample sizes cause issues due to the small number of items present within each assessment, the student sample size has the inverse issue, where the sample size is so large that any deviation from the model is seen to be statistically significant. This is because larger sample sizes have smaller error values, and thus a higher proportion of the values are likely to statistically significantly deviate from their expected values because of the small error that a large sample size generates. An example of the breakdown of the student ability measures can be seen in Table 13 (see Appendix 7.6 for the breakdown from all the assessments), which highlights the number of students undertaking each assessment.

Table 13: Exploratory Analysis of Student Ability Measures in Chemistry IA 2012 Assessments

	n	**Mean**	**Standard Deviation**	**Std. Error**
Lecture Test 1	469	0.508	1.251	0.058
Lecture Test 2	446	0.312	0.965	0.046
Exam Part 1	508	-0.208	1.162	0.052
Redeemable Exam	487	0.509	1.099	0.050

Based on the number of students present for each assessment it should be expected that the student distribution is apparent from a histogram, an example of which can be seen in Figure 10 (see Appendix 7.7 for the student measure histograms from all assessments analysed).

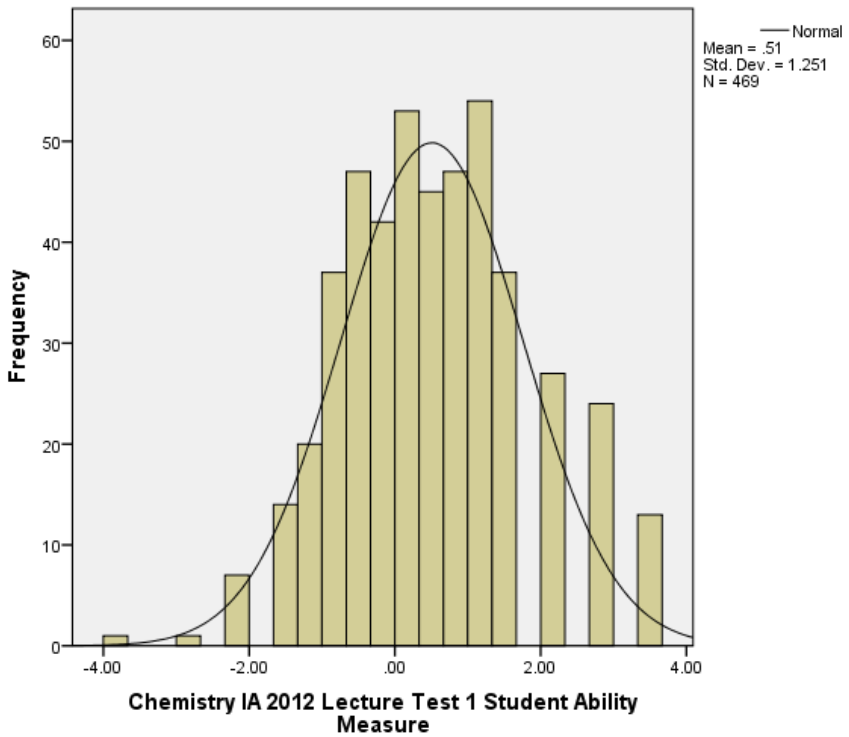

Figure 10: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IA 2012 to Determine the Distribution that the Measures Follow

The histogram shows that the student ability measures follow a roughly Normal distribution, however this can be better evaluated by reviewing the measures of spread within the distribution. The measures of spread for this task are reported within Table 14 (see Appendix 7.6 for the distribution analysis of the student measures from all assessments analysed), which provides a description for how the data is shifted from the Normal model.

Table 14: Measures of Spread of Rasch Student Ability Measures in Chemistry IA 2012 Assessments

	Skewness		**Kurtosis**	
	Value	**Std. Error**	**Value**	**Std. Error**
Lecture Test 1	0.340	0.113	0.222	0.225
Lecture Test 2	-0.054	0.116	-0.133	0.231
Exam Test	0.177	0.108	0.539	0.216
Redeemable Exam	0.402	0.111	0.033	0.221

Reviewing Table 14 it is clear that the distribution of the student ability measures within this set of assessment tasks being analysed follow the spread that is expected of a Normal model based on the values of skewness and kurtosis. There does appear to be a tendency for the ability measures to be slightly skewed higher than what is expected, but not to such an extent that it is a concern for the distribution following a Normal model. The tests of normality of the ability measure distribution within this set of assessment tasks being analysed can be seen below within Table 15 (see Appendix 7.6 for the tests of normality on all of the student ability measures generated from every assessment task analysed).

Table 15: Tests of normality of Rasch Student Ability Measures in Chemistry IA 2012 Assessments. Highlighted Cells indicate Observation of a Statistically Significant Difference

	Kolmogorov-Smirnov			**Shapiro-Wilk**		
	Statistic	**df**	***p*-value**	**Statistic**	**df**	***p*-value**
Lecture Test 1	0.086	469	<<0.001	0.976	469	<<0.001
Lecture Test 2	0.083	446	<<0.001	0.983	446	<<0.001
Exam Test	0.105	508	<<0.001	0.971	508	<<0.001
Redeemable Exam	0.070	487	<<0.001	0.984	487	<<0.001

The tests of normality show that on every occasion the distribution of the student ability measures are statistically significantly different from a normal distribution; however, the values of skewness and kurtosis suggest that the deviation is not large enough to cause issues with assuming a Normal distribution. It is likely that the large sample size is causing issues when attempting to test whether the data fits a Normal distribution, and therefore the measures of spread are a more reliable measure in this instance. Again, inspection of an example Q-Q plot shown in Figure 11 (see Appendix 7.7 for all Q-Q plots of student ability measures from the assessments analysed) clearly shows that there is not significant deviation of student ability from the Normal distribution except within the tails of the data, which is expected within an assessment task.

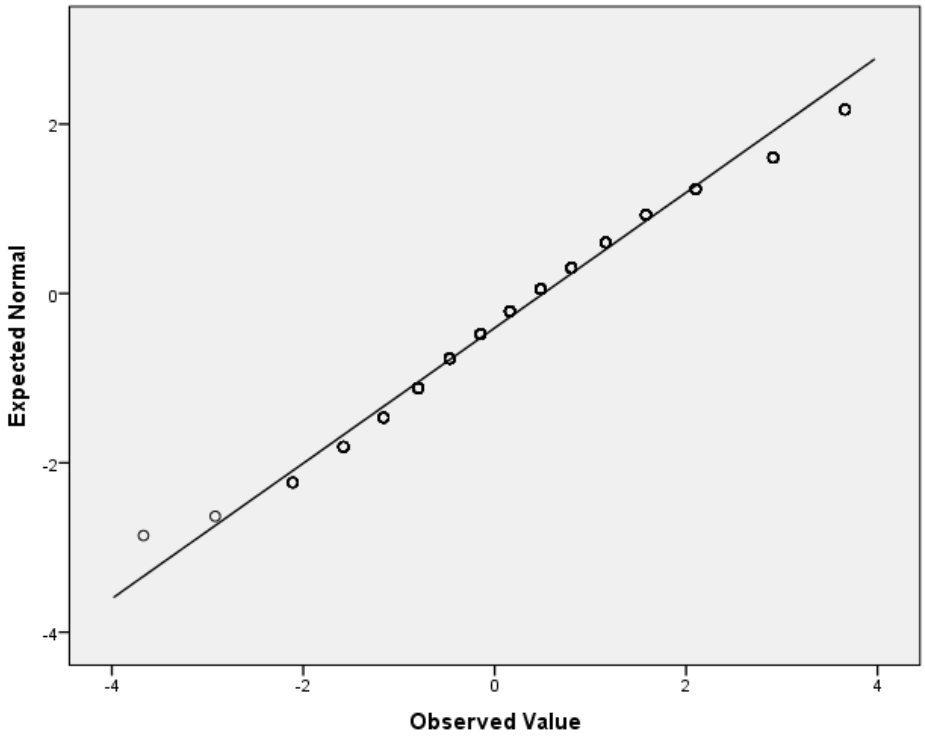

Figure 11: Q-Q Plot of the Rasch Student Ability Measures from Lecture Test 1 in Chemistry IA 2012

Analysing the other assessment tasks for their measures of spread, tests of normality, and Q-Q plots shows that even though the tests of normality almost always show a statistically significant difference from the Normal distribution (only on one occasion does an assessment task 'pass' the test of normality) the distribution of the data is not expected to significant deviate from a Normal distribution for the purpose of applying statistical tests. This is seen by the Q-Q plots only showing deviations within the tails of the data, and as only minor skewness deviations (skewness between -1 and -0.5 or between 0.5 and 1) are seen in seventeen of the sixty-four assessment tasks being analysed, this is thought to not represent a major issue in the distribution of the data. Therefore, there is no issue in applying statistical tests that assume a Normal distribution for the purposes of comparing student ability measures generated through the use of the Rasch model.

In this research, the ability and difficulty measures generated were used as a method for comparison and thus these factors needed to be explored even though fitting a Normal distribution is not required for evaluation of assessments applying Rasch analysis in most circumstances. If the intention was to analyse and improve each assessment somewhat independently then there is no need for the measures to follow a Normal distribution; for that analysis to be valid the critical measure is the fit statistics which inform how well the ability and difficulty measures match the expectations of the Rasch model.

3.3.3 Rasch Item Analysis

The most obvious misfitting items can be observed visually through the use of the Wright map and bubble charts. The Wright map can be used to identify items that provide less information than other items within the assessment task, as items that are either much too easy or much too hard for the cohort being assessed provide no effective information about the students. This would easily be observed as an item measure that lies significantly above or below the student ability measure distribution of the entire student cohort (shown on the right side of a Wright map as seen in Figure

6), or at the very least lies close to the extreme ends of the student cohort. Alternatively, a bubble chart can be used to visualise how far from the Rasch model each item deviates and the size of the error within that measure. As discussed within Section 2.5.3 the infit statistic is largely influenced by the results of students whose ability measure lies close to the difficulty measure of the items, and hence it is less likely to be influenced by outliers. Thus item infit is able to show clear fluctuations for the students that the item will obtain the most information about (i.e. the students who lie close to the measure); however, it does not account for the students who lie away from the item difficulty. The effect of the item on outlier students is captured by the item outfit statistic, which has no weighting to favour ability measures closer to the item difficulty measure. This means that outliers within the data can have a large impact on the outfit measure, and while it may sound problematic to have a measure largely influenced by outliers it is important to determine how the students at the extreme ends of the scale are responding to the item. Thus, outfit can be used to detect when students are clearly underperforming or over-performing on an item based on the comparison between their ability measure and the item's difficulty measure. For example, Figure 12 displays an infit bubble chart where each circle represents an item: the size of the circle is representative of its error range; its position on the x-axis is its infit measure; and its position on the y-axis is its difficulty measure. As each assessment item was to be numerically evaluated in this research, the bubble chart representation was not used extensively. It is presented here to demonstrate that it can be a highly effective method for obtaining a snapshot of how closely individual items match the model.

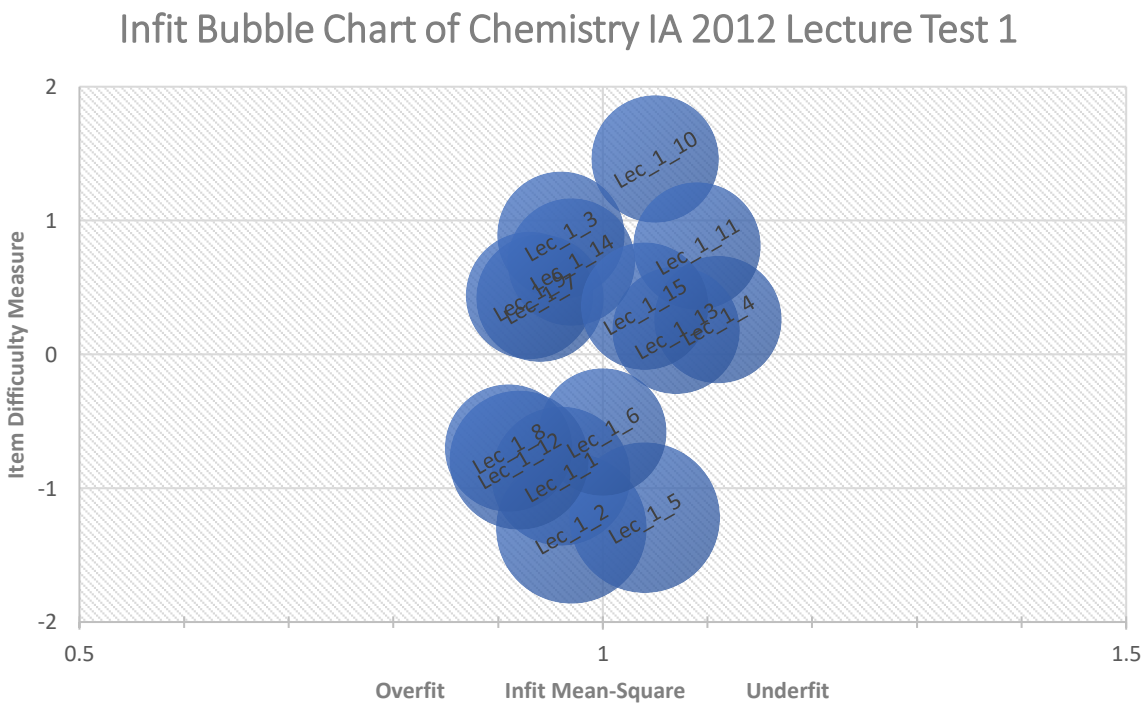

Figure 12: An Infit Bubble Chart of Lecture Test 1 from Chemistry IA 2012 to Visualise the Fit of the Items to the Rasch Model

It is also important that outfit is always analysed in the same way that infit is, and thus the bubble chart of the item outfit should also be analysed and can be seen below in Figure 13 for the same assessment task.

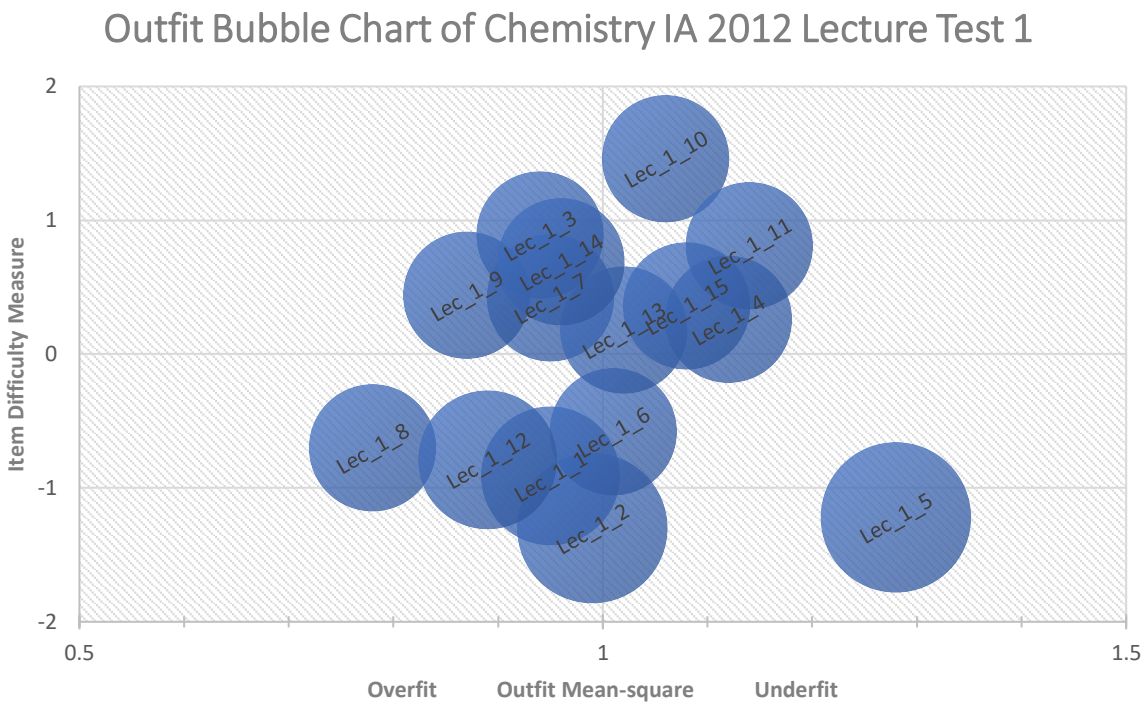

Figure 13: An Outfit Bubble Chart of Lecture Test 1 from Chemistry IA 2012 to Visualise the Fit of the Items to the Rasch Model

Generally, it is expected that for an item to appropriately fit the model the fit statistics (infit and outfit) should lie between 0.7 – 1.3; however, depending on the assessment task these guidelines can be changed to suit the purpose of the assessment.[273,309] Within this research, values of 0.8 – 1.2 (>1.2 implies more variance than expected [underfit], <0.8 implies more predictable than expected [underfit]) were used to define the deviation of infit and outfit that was allowed before an item was considered to be misfitting, as individual assessment items were utilized multiple times and thus the number of times an item was used should be considered to account for variance within results. Thus, items that were used multiple times but only showed significant deviation from the Rasch model once could be considered to be random variations, whereas if the item was found to be significantly different from the fit statistics on multiple occasions it suggests that there is an issue within the item itself rather than statistical variation causing it to appear significant.

Inspecting Figure 12, all the items used within the assessment task lie within the allocated infit range, implying that all the items reasonably fit the Rasch model. In contrast to this Figure 13 clearly shows items that appear outside of the desired outfit range suggesting problems within the items analysed, highlighting the importance of analysing both the infit and the outfit measures of an item. Visual inspection of Figure 13 suggests that item Lec_1_5 and item Lec_1_8 are worth further exploration, as are other items that are on the fringe of the desired range guidelines. As mentioned previously, these tools were not used extensively within this research; however, they are worth considering depending on the circumstances of the assessment and the analysis being undertaken. For example, on an assessment with a large number of items this could act as a first filter, or when performing continual assessment analysis these can be used as a quick tool to identify if any items have shown substantial shifts from how they have performed previously.

To evaluate how closely the items fit the model the measures of fit need to be evaluated numerically to determine if the fit of the item to the model is appropriate, and adequately explain the interactions between the students and the items. The measures of fit provided for the Rasch model for each item used in Lecture Test 1 from Chemistry IA (2012) are shown in Table 16 (see Appendix 7.5 for the fit measures for every item analysed). Knowing the actual item difficulty is important when evaluating if the individual item is performing as expected; however, the value itself is less important when analysing the metrics of its performance, as it does not give any information about the item's fit to the Rasch model. The standard error is useful both as a determinant of the confidence of the precision in the measures generated and can also be used to determine if two measures are statistically significantly different from each other. Generally, the two measures must be different by three times the standard error to be confident that the measures are significantly different from a statistical perspective. Both the observed versus expected values are important for determining that the item matches the rest of the assessment task and is the best indicator of issues unrelated to the assessment such as errors within option keying. The main measures that need to be evaluated are the infit and outfit values for all the items to determine if they show deviation from the Rasch model. Whether that variation is due to statistical variance or if it truly represents a statistically significant deviation from the Rasch model can be evaluated using the ZSTD measure. Using the knowledge obtained from the bubble chart it is known that item Lec_1_5 and item Lec_1_8 may be potential concerns for this task, which can then be evaluated by reviewing the item fit statistics as shown in Table 16.

Table 16: The Rasch Measures Calculated for Every Item in Lecture Test 1 from Chemistry IA 2012

	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS %	EXP %
Lec_1_1	-0.91	0.12	0.96	-0.70	0.95	-0.40	0.43	0.40	77.1	77.0
Lec_1_2	-1.30	0.13	0.97	-0.40	0.99	0.00	0.39	0.37	81.1	81.4
Lec_1_3	0.89	0.11	0.96	-1.00	0.94	-0.80	0.51	0.48	74.7	70.9
Lec_1_4	0.26	0.11	1.11	2.60	1.12	1.80	0.39	0.47	62.9	69.4
Lec_1_5	-1.22	0.13	1.04	0.60	1.28	2.00	0.32	0.38	80.4	80.4
Lec_1_6	-0.58	0.11	1.00	0.00	1.01	0.10	0.42	0.42	74.5	73.7
Lec_1_7	0.42	0.11	0.94	-1.30	0.95	-0.80	0.51	0.47	72.1	69.4
Lec_1_8	-0.70	0.11	0.91	-1.70	0.78	-2.30	0.49	0.42	76.5	74.8
Lec_1_9	0.44	0.11	0.93	-1.70	0.87	-2.10	0.53	0.47	71.6	69.4
Lec_1_10	1.46	0.11	1.05	1.00	1.06	0.60	0.45	0.49	72.7	75.2
Lec_1_11	0.81	0.11	1.09	2.10	1.14	2.10	0.41	0.48	67.9	70.3
Lec_1_12	-0.79	0.12	0.92	-1.50	0.89	-1.00	0.47	0.41	75.8	75.6
Lec_1_13	0.18	0.11	1.07	1.60	1.02	0.40	0.43	0.47	64.4	69.3
Lec_1_14	0.69	0.11	0.97	-0.60	0.96	-0.60	0.50	0.48	71.0	70.0
Lec_1_15	0.36	0.11	1.04	1.10	1.08	1.20	0.44	0.47	67.9	69.4

Based on the results of Table 16 it can be seen that both items Lec_1_5 and Lec_1_8 have outfit values (1.28 and 0.78 respectively) that place them outside of the pre-determined guidelines, and both values are known not to be the result of statistical variation based on the ZSTD value. As observed within the bubble chart, there are some items that show small deviations from the Rasch model based on their fit statistics; however, none of those items deviate significantly from the Rasch model and hence are not identified as causing any issues within the assessment task. By carrying out this evaluation on all the assessment tasks used, a list of items was prepared to display all the items

that misfit the Rasch model on more than one occasion (Appendix 7.9). There are 83 unique items that were identified to show misfit on multiple occasions (or only once if the item was only used on one occasion) out of the 261 unique items that were used in all of the assessments analysed, which represents a large percentage of the items being utilised. From these 83 items 33 of them represent only minor issues within the item (deviation outside the guidelines, commonly within the outfit measure, that does not appear repeatedly), 9 items are potentially major issues (large infit or outfit measures [<0.70, >1.30] on occasions, but are only observed as problematic less than half the time they are utilised), and 41 items represent major issues (consistent and large infit or outfit item measures [<0.70, >1.30 on at least one occasion]). Based on the number of items that are considered to be problematic by Rasch analysis it is likely that they have some influence on the outcomes of assessment tasks; however, that does not mean that the assessment tasks are expected to be invalid. This is because the problematic items are spread across all of the assessment tasks, which means that it is unlikely that there are enough problematic items within an assessment task to invalid all of the results, and if there were it should be more evident within the measures obtained for each assessment task. Even though the results of the assessment tasks are not invalidated by these problematic items they do influence the ability of the assessment task to complete their purpose and are still a threat to the assessment's validity. Thus, it is critically important that any item that is identified, even if it is thought to only be a minor issue, is investigated further to determine the root cause of the issue so that it can be addressed.

The two different methodologies employed within this research to identify problematic items within the MCQ assessments showed a large amount of overlap when analysing the assessment task; however, they differ substantially when analysing the individual items. Classical Test Theory identified 12 unique items that were consistently problematic across all of the assessment tasks being analysed (4 major, 4 potentially major, and 4 minor), compared to Rasch analysis which identified 83 unique items that were consistently problematic (41 major, 9 potentially major, and 33 minor). There were only two items identified as problematic by CTT that were not also identified as problematic by Rasch analysis, and both these items represent minor issues based on CTT analysis. These large differences in the results obtained between the two methodologies can be attributed to the differences in their assumptions and the harsher criteria that can be used by the Rasch model due to those assumptions. One crucial difference is that CTT will never identify overfit items as an issue within an assessment task as they will not lie outside the thresholds used for identifying problematic items. These differences between the two methodologies are the cause of the disparities between the items identified as problematic, which is why it is important to consider the purpose of the analysis before an analytical methodology is selected. All but two of the items that CTT did identify matched underfit items identified by Rasch analysis, and thus CTT is performing as it is expected to when its assumptions are accounted for. While the results of CTT were considered within this research, because all of the information that it provided was also given within Rasch analysis (aside from the two minor issues identified), it meant that the results of Rasch analysis were used to inform the analysis of the item construction.

3.3.4 Breaking Down Item Construction

Identifying the items that may be causing problems within assessments is only part of the solution, as after identifying the items themselves, what to do with them needs to be considered. They could be removed from the assessment to stop them from influencing future results; however, if they are removed then they need to be replaced by new items within the assessment and there is no guarantee that the new items would be any better than the items being removed from the assessment. The alternative to this is to improve the items that are causing problems within the

assessment to stop them from causing issues within future assessments. The problem with this approach is that just because it is known that the item is problematic does not mean that the source of the problem within the item can be easily identified and corrected. It is entirely possible that what is causing the issue within an item may be so deeply embedded within the item itself that fixing the item is harder than creating an entirely new item. To determine where the root of the issue is the two main facets of a MCQ need to be broken down and analysed: the stem and the options.

The stem of an item should inform the students of the topic that the item is based around and should give them the actual question that can be answered. Ideally, the question posed within the stem of an item should be answerable without requiring the students to check the options presented to them.[43,45,75,77,219,221,222] Neither CTT nor Rasch analysis provide any numerical measures to determine if it is the stem that is causing the issue, and thus the only way to check the stem is to evaluate the stem of each of the items that were determined to be problematic. Evaluation of the stem can also help to identify if the issues identified by CTT or Rasch analysis make sense based on what the construction of the item requires from the students; for example, the stem of one of the overfit items identified (Chem IA, 2012, Lec_1_1: Problematic Fit Measures - Outfit: 0.79, 0.75 Infit: 0.92, 0.90 [Minor]) reads:

Sodium emits light of wavelength 690 nm when heated in a flame. What is the frequency of this light?

An overfit item means that the low ability students struggle more than expected while once the student reach a certain ability level they almost always obtain the correct answer. In this example all that is required of the students is to input the number given into the correct equation that is provided on their equation sheet and calculate the result. The only information that the students can gather from the options in this item is whether the value calculated is one of the options present, and if it is not then they know that they have made a mistake. Thus it should be expected that higher ability students have no issue with calculating the answer; however, lower ability students may be focusing on the fact that the light comes from sodium, or they may not be able to identify the correct equation. Thus, it is reasonable that considering this item was only significantly misfitting the Rasch model 2 times it was asked out of 8 occasions it was used within an assessment task, and that the potential overfit is acceptable within the Rasch model (only marginal deviations were seen within the outfit [0.79 and 0.75] and the infit values were within a reasonable range though tended towards overfit [0.92 and 0.90]), that no changes need to be made to this item. If the fact that low ability students are having a harder time than expected with this item is a concern for the assessors then it would be more effective to spend more time covering the concepts within lectures or tutorials than replacing the item.

Depending on the question being asked of the students, it may not always be possible for the item to be answerable from the stem alone. This is particularly true in items where the students need to either evaluate the options presented to them, or compare between them. Even if the item cannot be answered by the stem it should still be providing the context for the item, and getting the students thinking about the relevant concepts. The difference can be seen through the comparison of the following two problematic item stems (Chem IA, 2012, Lec_2_15: Problematic Fit Measures - Outfit: 0.70, 0.64, 0.66, 0.68 Infit: 0.83, 0.80 0.80, 0.85 [Major]) (Chem IA, 2012, Lec_2_6: Problematic Fit Measures - Outfit: 1.25, 1.24, 1.26 Infit: 1.19, 1.20, 1.17 [Minor]):

Which of the following species can function as a chelating ligand?

Which of the following is false?

The first stem gives the students a clear and well-defined question that gives them the context of the problem being presented to them, and even though the students are likely to evaluate each option it is possible that they may identify the correct option without comparing the options. The second stem gives the students no context, and all the students can take away from the stem is that they need to evaluate every option and choose the one that is incorrect. There are several issues that arise from this lack of information; the first is that the students are completely unaware of what the item is assessing them on, potentially even after they read the item's options. Another potential issue is the use of a negative within a stem - getting the students to find which option is false can lead to issues that are unrelated to student ability.[45,133] This is a combination of the stem lacking useful information and the students being more familiar with selecting the options that represent true statements, which may result in the students ignoring the negative within the stem and instead select one of the options they know to be true. There are two simple changes that can be made to help resolve these issues: providing context and highlighting the negative.

Original: *Which of the following is false?*

Adjusted: *Which of the following statements relating to trends within the periodic table is* ***FALSE****?*

There is no guarantee that these changes will lead to improvements in the performance of the item; however, the changes do directly address the concerns held about the stem and the item should be monitored and evaluated in future uses within assessments. The last common issue that was found within problematic items is that some stems do not include all the information required to answer the item. This often occurs when the options need to be evaluated in some way, and thus the students need to read the stem to gather the options and then evaluate them based on what was asked of them in the stem. Some examples of the problematic items that included this style of stem are shown below (Chem IA, 2012, Lec_1_7: Problematic Fit Measures - Outfit: 0.80, 0.77, 0.72 Infit: 0.86, 0.83, 0.83 [Minor]) (Chem IA, 2012, Lec_2_3: Problematic Fit Measures - Outfit: 1.37 Infit: 1.10 [Major]):

Making use of the Molecular Orbital energy diagram presented above, which molecule is paramagnetic?

Arrange the following atoms in order of increasing radius:

In both of these cases the stem asks a clear question of the students; however, the students are unable to provide an answer to the question without knowing the options that they need to evaluate. The first stem is completely reasonable as listing the options within the stem and then again as answer options would include information needlessly; however, more emphasis should be placed on the fact that the molecules in question are presented to the students within the options. The stem of the second item shown here does not give the atoms that need to be arranged according to size and instead the item goes directly to the options, which are presented as ordered lists. Additionally, not all of the options listed are atoms as some of them are ions. It could easily be argued that it is no more difficult for the student to extract the atoms they are required to order from the options presented within the item than it is for them to identify the molecules that need to

analysed using a molecular orbital energy level diagram; however, the important difference between the two stems is that the atoms that need to be ordered are not clearly presented within the options whereas the atoms that need to be identified as paramagnetic are. An item should be constructed to minimise the amount of work that is required of the students outside of answering the item, and thus it is completely reasonable to list the atoms within the stem itself to make it clear which atoms need to be arranged. A simple change to the new stem shown below requires next to no effort for the assessors and ensures that there can be no confusion from the students as to what atoms to arrange.

Original: *Arrange the following atoms in order of increasing radius:*

Adjusted: *Arrange the following species in order of increasing atomic radius: Si, S, F^-, Mg^{2+}*

Breaking down the stem of an item is an important step in determining where the students may be having issues, but even when the stem can be improved it does not mean that it is the root cause of an issue. This can be seen by evaluating the item stem below from item 9 in lecture test 2 from Chemistry IA 2012 (Problematic Fit Measures - Outfit: 1.81, 2.70, 2.03, 2.41 Infit: 1.23, 1.34, 1.21, 1.32 [Major]):

Phosphoric acid (H_3PO_4) can lose three protons to form phosphate ion (PO_4^{3-}), but phosphonic acid (H_3PO_3) can only lose two protons to form phosphite ion (HPO_3^{2-}). This is because

- *(A) P is in oxidation state +5 in H_3PO_4 but in oxidation state +3 in H_3PO_3*
- *(B) H_3PO_4 has three –OH groups and one terminal oxygen atom*
- *(C) H atoms in –OH groups are acidic and ionize to give H^+ ions*
- *(D) H_3PO_3 has one H atom bonded to P and one terminal oxygen atom*
- *(E) H_3PO_4 has three –OH groups and H_3PO_3 has only two.*

The key issue with this stem is that it does not ask a question of the students, but rather uses a "complete-the-sentence" style of stem, which means that the stem is unfocused and leads the students to spend more of their time evaluating the options rather than focusing on the question being asked.[45,75] This can be easily fixed by replacing "this is because" at the end of the stem with "why?" instead, as it changes the stem to be asking a question of the students. While that is an issue within the stem, it is unlikely that it is what was causing the item to be problematic all four times that it was used within assessments, based on the size of the issues observed (major issues in both infit and outfit on all four occasions). Thus, the answer options also need to be evaluated to determine what other potential issues are present. There are three different methodologies to analysing the options that can be used in conjunction with each other to determine what may be causing the issue. The first is to break down the option construction in a similar way to breaking down the stem; the second is to see which options the students favour at different ability levels using an item characteristic curve; and the third is to numerically evaluate how often each option is being selected and which options are performing in unexpected ways through the use of distractor analysis. The first step is to review the options themselves to determine if there are any glaring issues with the way that they are written and presented to the students.

The correct option for this item is option D, and by simply reading all the options as they are presented there does not seem to be any glaring issues with how they are written. This is where knowledge of the topic being reviewed is important, as even though the actual construction of the options is reasonable, the problem with them is how students are interpreting what the correct

option should be. This can be identified from the item characteristic curve (ICC). Before discussing the ICC for this specific item, it is worth considering a 'non-problematic' item. It is expected that the correct option follows a logistic curve and the distractor options should fall off around it. This is shown for item 9 in lecture test 1 from Chemistry IA in 2012 within Figure 14.

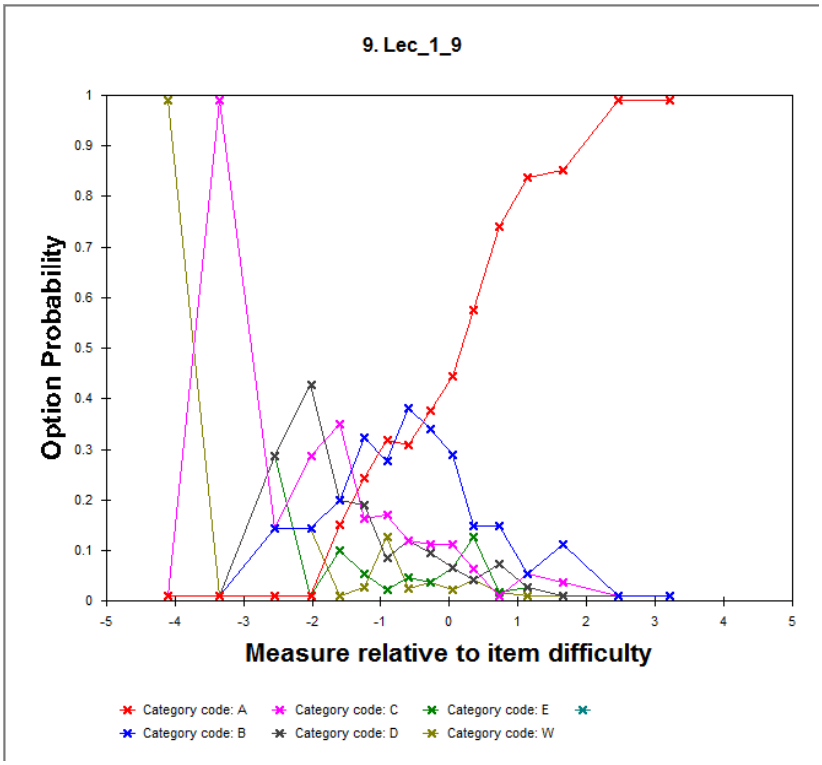

Figure 14: Option Item Characteristic Curve for Item_9 in Lecture Test 1 from Chemistry IA 2012 Displaying the Student Ability – Item Difficulty against the Probability of Selecting Each Option

The expected ICC shown in Figure 14 clearly shows the probabilistic nature that is expected from the Rasch model, as the higher the student ability is relative to the difficulty of the item (as displayed on the x-axis by "student ability" minus "item difficulty") the greater the probability that they select the correct answer (Option A - represented in red). Therefore, as the student ability increases the probability that they select distractor options decreases, which is observed through the decrease of the selection rates of the distractors (Options B – E, where category W represents students leaving the response blank – each of which is represented by a different colour within the graph). It is expected that all ICCs will show the selection rates of the distractor options decreasing, and the selection rates of the correct option increasing as the student ability becomes higher relative to the item difficulty. Any item whose ICC does not match this trend does not follow the fourth assumption of the Rasch model (the difference between student ability and item difficulty, and nothing else, can be used to predict the probability of observing any scored response), and therefore is likely to be determined to be a misfitting item based on the expectations of the Rasch model.

The ICC for the problematic item (item 9 in lecture test 2 from Chemistry IA 2012) is shown in Figure 15 below. There are three important considerations that can be seen by reviewing Figure 15: the first is that the correct option, (option (D) (the black line), is not following a smooth logistic curve as expected, and the option that might otherwise be expected to represent the correct option (option (E); the green line) is actually one of the distractors. This is an immediate sign that there is an issue

within the item, particularly in relation to option (E) as the higher ability students select this option more often.

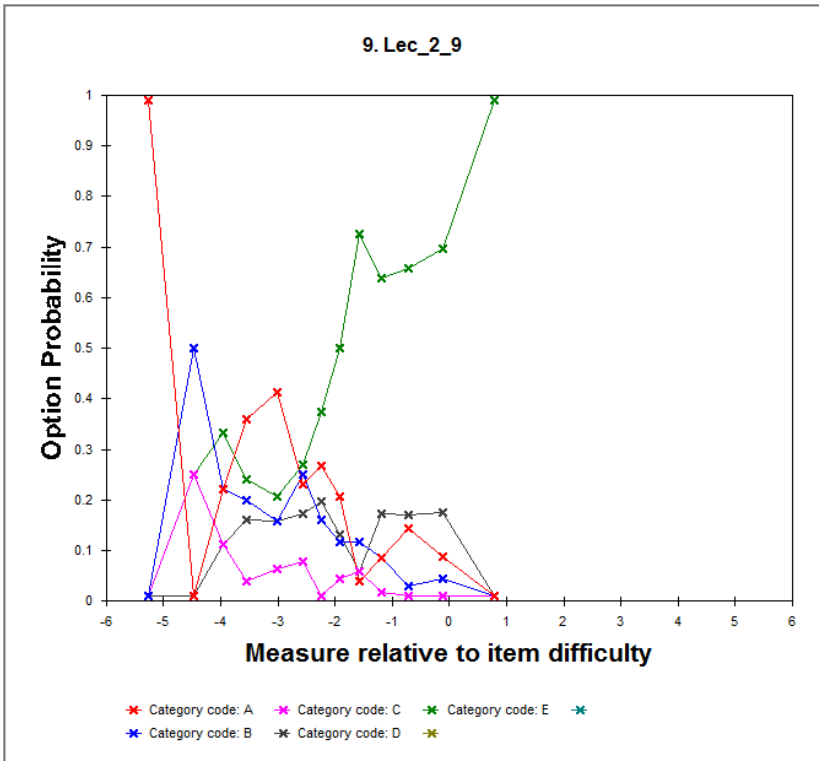

Figure 15: Option Item Characteristic Curve for Item_9 in Lecture Test 2 from Chemistry IA 2012 Displaying the Student Ability – Item Difficulty against the Probability of Selecting Each Option

The second notable factor is that the graph appears to be significantly shifted to the left compared to the 'non-problematic' example ICC shown in Figure 14, which is because the x-axis is the relative measure of student ability compared to item difficulty (i.e. student ability – item difficulty) and means that this item has a high item difficulty measure (in this example the item difficulty was 2.34, making it quite high relative to the other items within that assessment task). The last consideration is the behaviour of the distractor options, as it is expected that once the options begin to be selected less by higher ability students, they should show continuous decrease in selection until they are no longer selected, as seen within Figure 15. That is not the case with the distractors present within this item as they continue to show spikes and dips throughout almost all of the different student ability measures. Visually that gives a lot of information about the student selection behaviour and what options they trend towards; however, that information can be quantified to provide more accurate information about the performance of each option. This information can be seen in Table 17 and presents itself like an item level analysis for each individual option.

Table 17: Distractor Analysis of Item 9 in Lecture Test 2 from Chemistry IA 2012

Option	Value	Count	%	Ability Mean	P. SD.	S. E. Mean	Infit	Outfit	Point Measure Correlation
A	0	93	21	-0.11	0.89	0.09	0.8	0.7	-0.23
B	0	62	14	-0.10	0.88	0.11	0.7	0.7	-0.17
C	0	18	4	-0.22	0.87	0.21	0.6	0.6	-0.11
D	1	67	15	0.34*	0.93	0.11	1.7	1.9	0.01
E	0	206	46	0.67	0.89	0.06	1.3	1.5	0.34

The distractor analysis shows the amount of marks that each option is assigned (value), the number of students that selected each option (count), the percentage of the student cohort that selected that option (%), the mean student ability measure of the students that selected the option (ability mean), the population standard deviation of the ability measures for each option selected (P. SD.), the standard error in the ability mean of each selected option (S. E. Mean), the infit value for that option based on the students who selected it (infit), the outfit value for the option based on the students who selected it (outfit), and the correlation between the value assigned to the option and the students' ability measures. All of this information can be used to determine what sort of students are selecting each option, and what that implies about each option's functionality within the item.

The most important measure when evaluating the options in an item is the correlation, as it is expected that all of the distractor options show a negative correlation to the final result whereas the correct option should be positively correlated to it. This is because students that select the incorrect answer should be expected to perform less well than students who select the correct answer, as otherwise it indicates that what the item is assessing the students on is something unrelated to the rest of the assessment task. In this case both the correct option (D) and one of the distractors (E) are positively correlated, with option (E) being more highly correlated than option (D), as option (E) is selected by higher ability students on average (as highlighted by the asterisk next the ability mean of the correct answer). Even though option (E) is positively correlated the size of the correlation is close to 0, thus implying that student success on this item is not reflective of their performance across the rest of the assessment task. Immediately this highlights that there is either a problem with how option (E) is presented or how option (D) is presented, or potentially a combination of the two. A quick review of the other distractors shows that option (C) is considered dysfunctional (has < 5% student selection; it is unsurprising to have at least one dysfunctional distractor in a five option item, as generating four functional distractors for each item is not always possible and it is likely that at least one of them will be a weaker option that is dismissed by the students). Option (A) is selected more often than the correct option, and option (B) is very close to the same selection rate; this is not necessarily an issue for a difficult item but it is worth considering how difficult the assessors expect this item to be in relation to how difficult the item is found to be by the students. Knowing how the options are performing within the assessment, they can be reviewed to determine if these results are expected or if the behaviour of the students is completely unexpected.

If each of the options is considered based on the information gained from distractor analysis, decisions can be made about what options may need to change. Option (C) is known to be dysfunctional, which is likely as it clearly does not address the topic being discussed (even though it is a true statement) and hence is seen by the students to be irrelevant to the item and thus dismissed as a potential answer option. Option (B) sees almost the same amount of selection as the correct option, which makes sense as it is almost a mirror of the information presented within the correct option. The correct option addresses only one of the two molecules and informs the reasoning for why that molecule can only lose two protons, whereas option (B) informs the reasoning as to why the other molecule can lose three protons. Thus, it is reasonable that the students are just as likely to be selecting option (B) as they are the correct answer, and hence if the answer option is not changed then this distractor should be modified. Option (A) addresses both molecules, but it clearly does not address the topic being discussed and thus is simply a highly effective distractor within this item and there is no reason to change it. Option (E) is clearly favoured by high ability students, likely because they believe it gives a more complete answer to the problem

than the correct answer option does. It is likely that this is the case because unlike the correct answer it addresses both of the molecules discussed within the stem, and with some chemistry knowledge (the information provided in option C) the importance of that difference does answer the item. However, the assessors seem to want the students to focus specifically on explaining only the properties of H_3PO_3, and that is why option (D) is the correct option as it explains why the third proton in the molecule cannot ionise. It could be argued that none of these options provide a clear answer to what was being discussed within the stem, and in fact considering all of the options together provides a much clearer understanding as to the differences between the two molecules. This is why the MCQ format is called the single-best response, as multiple options can be true, but students need to pick the most correct one; however, in this instance the students and the assessors disagreed over what option that was. Thus, to improve upon this item, options B, D, and E need to be considered to be changed to make it clearer which option provides the single-best response. It may also be worthwhile to highlight within the stem that the item is focusing on H_3PO_3 specifically, and thus responses need to relate to that molecule to avoid student confusion. The importance of how students approach items is highlighted within this item, as different student approaches may change the option that they select and thus influence the problematic nature of the item. This particular item was actually changed within the years of assessments covered within this research, and as a result, half the times it was asked it was as the problematic item discussed and the other half it had been changed, as outlined below:

Original Item:

Phosphoric acid (H_3PO_4) can lose three protons to form phosphate ion (PO_4^{3-}), but phosphonic acid (H_3PO_3) can only lose two protons to form phosphite ion (HPO_3^{2-}). This is because

- *(A) P is in oxidation state +5 in H_3PO_4 but in oxidation state +3 in H_3PO_3*
- *(B) H_3PO_4 has three –OH groups and one terminal oxygen atom*
- *(C) H atoms in –OH groups are acidic and ionize to give H^+ ions*
- *(D) H_3PO_3 has one H atom bonded to P and one terminal oxygen atom*
- *(E) H_3PO_4 has three –OH groups and H_3PO_3 has only two.*

Revised Item:

Phosphoric acid (H_3PO_4) can lose three protons to form the phosphate ion (PO_4^{3-}), but phosphonic acid (H_3PO_3, also known as phosphorous acid) can only lose two protons to form the phosphite ion (HPO_3^{2-}). This is because

- *(A) P is in oxidation state +5 in H_3PO_4 but in oxidation state +3 in H_3PO_3.*
- *(B) H_3PO_3 has two P=O groups.*
- *(C) H atoms in –OH groups are acidic and ionize to give H^+ ions.*
- *(D) H_3PO_3 has one H atom bonded to P and one terminal oxygen atom.*
- *(E) P is more electronegative than O.*

Within the stem of the item more information about phosphorous acid was included, but nothing that would be expected to significantly alter the students' perception of the item. Both option (B) and option (E) have been completely changed from the distractors that they were previously to ensure that students can clearly identify option (D) as the single-best response. There was no emphasis placed on focusing on only H_3PO_3 to avoid student confusion, or changes to the answer to address both molecules, but despite this the changes to the distractors were able to rectify the item and prevent it from being problematic. This can be seen within Figure 16, which shows the distractor item characteristic curve for the new version of the item.

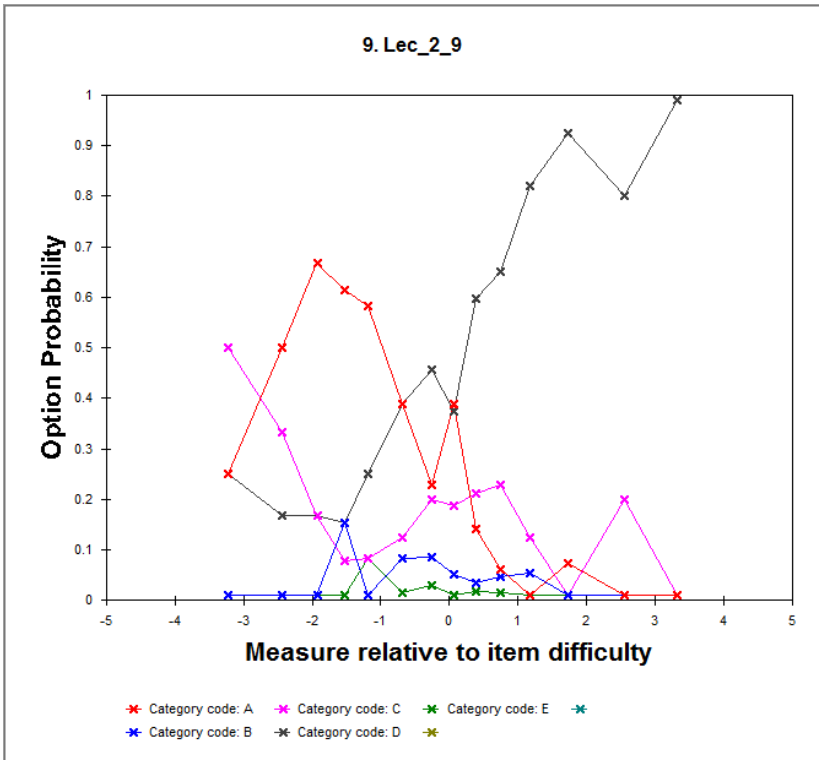

Figure 16: Option Item Characteristic Curve for Item_9 in Lecture Test 2 from Chemistry IA 2014 Displaying the Student Ability – Item Difficulty against the Probability of Selecting Each Option

Looking at the figure it can be seen that that the correct option now follows much more closely to the logistic curve that is expected from it than it previously did, and that the graph has significantly shifted toward the right due to the change in the students' results. This is a result of changes to the item significantly lowering the item difficulty (previously item difficulty was 2.34, the new item has a difficulty of 0.37) as students are no longer confused by the options presented to them, which displays the importance of evaluating and improving items after they have been used in assessments. A significant lowering of item difficulty based on changing only two distractor options suggests that the item was flawed and causing issues that were unrelated to what it was supposed to be assessing, and thus by identifying the issue and changing it, the item was improved for future assessments.

3.4 Applications of Item Analysis

3.4.1 Using Item Analysis to Identify Gender Differences and Categorise Items

What else can be learnt from item analysis outside of determining if the items are performing as expected of them within the assessment task should be considered when performing an item analysis, as the purpose of the analysis may not be addressed by determining which items are problematic. Within this research there were two other ways in which the items were analysed to determine the impacts of their performance and their function within an assessment task. These other modes of analysis were comparing the male and female student cohorts for differences between how they perform on specific items, and categorising the items based on their construction.

Any item favouring either males or females and resulting in a statistically significant difference between the results obtained of these two groups is a threat to the validity and the purpose an

assessment task, as it is giving one group of students an advantage over a different group. If these items continue to be used within the assessment it means that the results of the students may not reflect their ability, which may influence their future decisions and pathways available to them. Thus, it is important that if there are any items that behave in that manner, they are identified so that steps can be taken to remove the issue from the assessment task.

If the male and female student cohorts perform statistically significantly differently on an item, then for that particular item one of the two cohorts must either find that item more or less difficult than the other cohort. If both the male and female student cohorts share the same distribution of ability, then it should be expected that male and female students have the same probability of selecting the correct option within an MCQ assessment. It is possible to compare the male and female student cohorts based on their results within the assessment task to determine if the cohorts share the same distributions of ability; however, if there is a large proportion of items that favours one gender then it is possible that they may skew the results and influence the outcomes of the assessment task. Therefore, while the cohorts can be compared using the results of the assessment task, when the items are analysed individually it should be assumed that there is no difference between the male and female student cohorts and the results of the cohort comparison can be considered after the item analysis has taken place. For example, if it is determined that the male student cohort has a statistically significantly higher ability level than the female student cohort then while it should be expected that all of the items will reject the null hypothesis that male and female students both have an equal probability of selecting the correct answer (as male students should be expected to be more likely to due to their higher ability), that null hypothesis still needs to be used. There are a couple of reasons for this, the most important of which is to ensure that when there is a difference between the two cohorts observed that it is also reflected within the items. If the cohort difference is not reflected within the items then it either means that the particular item being analysed may favour the lower ability cohort, or it may imply that the results of the cohort comparison have been skewed by gender biased items. Another reason to keep the same null hypothesis even when it may not be expected to be true is that it provides results relative to the problem being addressed. When attempting to identify gender bias, knowing exactly how much bias the item contains is secondary to determining if there is bias present, and therefore there is no reason to use a null hypothesis that complicates the results of the analysis. Therefore, within this research the student cohorts were compared before the individual items were analysed, but no changes were made to the item analysis based on the results of the cohort comparison.

The other item analysis that was undertaken within this research was classifying the items based upon their construction and how they are expected to be answered within the assessment. The purpose of item categorisation is to be able to describe an item and how it functions, which can be used to inform the expectations placed upon the item, ensure that the item matches the purpose of the assessment, or to help generate other items that are similar. Viewing the categorisation of all of the items within an assessment task can also be used to help ensure that the task is constructed in such a way that it matches its purpose and each of the items contributes towards that. Categorisation can be used in this manner before the assessment task has taken place, or it could be used after the assessment task has been administered in conjunction with item analysis to determine if there are any trends with the items that are causing problems within the assessment. If any trends are identified in this manner they can be applied in future assessment analyses to help identify specific items that have the potential to cause issues within the assessment. It is also possible to use the item categorisation to generate a new item that closely resembles an old item that may either be outdated or identified as problematic in some way and needs to be replaced. By

ensuring that the new item is categorised similarly to the previous one it allows for the item to be replaced without changing the overall construction of the assessment task outside of a few small adjustments that may need to be made to the new item.

3.4.2 Considerations of Student Ability and Differences in Gender Performance

Before discussing the potential of whether there is a difference in how male and female students answer specific items it is important to consider if there is expected to be a difference in how the two cohorts perform within the assessment. If the ability level of one of the cohorts is statistically significantly higher than the other then the differences between how they perform on an item is potentially a result of the ability level difference, and not due to any issues within the item itself. It would be expected that there would not be a statistically significant difference in ability level of the two different cohorts; however, there is the potential that due to self-selection and the amount of choice that is available to students when studying at a tertiary level that this could cause a shift in the ability levels of the cohorts. Determining if there is a statistically significant difference between the student cohorts will rely on using the raw scores obtained by the students within the assessment task if using a CTT approach, and the student ability measures when using Rasch analysis.

An issue that needs to be considered within the results of the assessment tasks being analysed whenever either methodology is used in this research specifically is that the students have the opportunity to use their best result from the combined lecture tests or the redeemable section within the final exam, which may influence the students' behaviour toward the assessments. Having the safety net of the redeemable assessment task means that students may approach either of those tasks in ways that they would not otherwise do, such as not studying beforehand, attempting to learn the assessment, and treating the task as a practice test. There is no way to account for this sort of behaviour by the students, but it is reasonable to assume that the behaviour is not isolated to either male or female students, and thus it is not expected to result in statistically significant differences between the two cohorts. This does mean that within both CTT and Rasch analysis the minimum scores (i.e. students who scored 0 on the assessment) had to be ignored, as there is no way of knowing if this result represented the student obtaining the minimum score, or if this was the result of some other factor such as a particular approach toward the assessment task. Another consideration when undertaking this analysis is that some students will not answer all of the items presented to them, which is unexpected as within these assessment tasks there is no negative penalty for giving an incorrect answer, and given that it is an MCQ assessment it would be expected that the students would at least provide an educated guess. This does not pose an issue within Rasch analysis, as the student ability calculations can account for students not providing answers on items; however, as CTT uses the raw scores it assumes that the students all attempted the same number of items. It cannot be known why students leave specific items blank; it may have been deliberately done by the students acknowledging that they do not know the correct answer, or it is possible that the students never attempted the item due to the redeemable nature of the assessments. Arguably the students not answering items should be considered to be the same as giving an incorrect answer, as they are expected to provide answers to every item within the time frame provided. For most assessments this would be the case; however, due to the redeemable nature of the assessment it cannot be known exactly why the students are leaving answers blank, but it should not be assumed that this is the same as incorrectly answering the item, even though it is treated as such.

Using CTT male and female students can be compared based purely on their raw score using an independent sample t-test. The results for the student cohort from one of the first-year courses at

The University of Adelaide can be seen in Table 18, which shows that in 7 of the 16 assessments male students perform statistically significantly better than female students.

Table 18: Comparison of Male and Female Student Cohorts undertaking Chemistry IA using the Student Raw Scores to determine if Student Ability is Significantly Different between them. Highlighted Cells indicate Observation of a Statistically Significant Difference

Chemistry IA	2012		2013		2014		2015	
	d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value
Lecture Test 1	466	0.006	445	0.223	470	0.266	502	0.020
Lecture Test 2	444	0.175	418	0.006	434	0.029	449	<<0.001
Exam Part 1	499	0.010	499	0.319	505	0.212	538	0.377
Redeemable Exam	485	0.977	486	0.546	496	0.021	523	0.237

While the number of assessment tasks that display statistically significant differences indicates that in most cases the ability of the male student cohort is higher than the ability of the female student cohort there are a few factors that need to be considered. The first consideration is that the size of the student cohort being analysed is quite high, and thus it is possible that even small differences between the cohorts result in statistically significance being observed. This is because when sample sizes are large their error becomes smaller due to the amount of information provided within the sample, and therefore it becomes easier for other measures to be considered to deviate statistically significantly from the large cohort.[269] Another consideration is that the course that is being analysed is a first semester course that has the prerequisites of Stage II Chemistry, and thus student self-selection may influence the type of students enrolling in the course at this stage. Self-selection represents the students choosing their own pathways, and thus different types of students will enrol in different courses based upon their motivations and goals. Chemistry IA (and Chemistry IB) requires the students pass SACE Stage II Chemistry with a C+ or better, and thus the students had to actively take these courses before coming to university demonstrating some forethought in their actions. However, there is a large range of abilities that may still be accepted into Chemistry IA based on that prerequisite, students that achieve a C+ compared to those that achieve an A+ likely have a large gap in ability. A student's motivation and goals do not need to directly align to cause them to enrol within Chemistry IA, as some students may be required to enrol as part of their program of study. These students may not be motivated to perform in the course as even though it is a requirement of their program, they may simply wish to pass the course to move forward in their degree. These are all factors and considerations in why a difference may be observed in the results of the male and female student cohort. The other first-year course with prerequisites takes place in the second semester, after the students undertake another round of self-selection based on their experiences at university, and such significant differences are not seen between male and female student cohorts in that course, as shown below in Table 19.

Table 19: Comparison of Male and Female Student Cohorts undertaking Chemistry IB using the Student Raw Scores to determine if Student Ability is Significantly Different between them. Highlighted Cells indicate Observation of a Statistically Significant Difference

Chemistry IB	2012		2013		2014		2015	
	d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value
Lecture Test 1	380	0.203	376	0.072	421	0.972	426	0.002
Lecture Test 2	361	0.595	346	0.010	392	0.060	389	0.005
Exam	425	0.847	446	0.879	479	0.070	478	0.509
Redeemable Exam	419	0.567	432	0.810	454	0.511	469	0.706

It can be seen from the table that only 3 of the 16 assessment tasks analysed show statistically significant differences between the male and female student cohorts within Chemistry IB, where all of the significant differences show the male cohort achieving higher results. The difference between Chemistry IA and Chemistry IB is likely due to the additional period of self-selection after the students have had the experience with a first-year chemistry course and can determine if it is something that they wish to continue studying. The two courses that do not have prerequisites, and do not show the same issues as seen within the other courses, can be observed within Table 20 and Table 21.

Table 20: Comparison of Male and Female Student Cohorts undertaking Foundations of Chemistry IA using the Student Raw Scores to determine if Student Ability is Significantly Different between them. Highlighted Cells indicate Observation of a Statistically Significant Difference

Foundations of Chemistry IA	2012		2013		2014		2015	
	d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value
Lecture Test 1	257	0.925	307	0.902	250	0.922	292	0.223
Lecture Test 2	265	0.132	253	0.723	221	0.473	234	0.693
Exam Part 1	301	0.315	360	0.611	323	0.211	363	0.757
Redeemable Exam	256	0.926	334	0.157	299	0.251	329	0.539

Table 21: Comparison of Male and Female Student Cohorts undertaking Foundations of Chemistry IB using the Student Raw Scores to determine if Student Ability is Significantly Different between them. Highlighted Cells indicate Observation of a Statistically Significant Difference

Foundations of Chemistry IB	2012		2013		2014		2015	
	d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value
Lecture Test 1	234	0.604	247	0.935	214	0.961	229	0.036
Lecture Test 2	187	0.042	216	0.864	182	0.947	196	0.304
Exam	264	0.125	296	0.746	274	0.694	297	0.075
Redeemable Exam	248	0.376	286	0.204	257	0.864	275	0.290

There is no statistically significant ability difference between the male and female student cohorts within Foundations of Chemistry IA; however, there are two instances where statistically significant differences were observed within Foundations of Chemistry IB. In contrast to what was observed within the prerequisite courses the statistically significant difference observed in the 2012 assessment task showed that the female student cohort statistically significantly outperformed the male student cohort. The other statistically significant difference observed within 2015 favoured the male student cohort, which suggests that differences between the male and female student cohorts are less likely to occur within the non-prerequisite courses. The lack of prerequisites likely mean that

the majority of the students are perhaps more likely to have only a limited amount of chemistry knowledge prior to starting the course, and thus student selection of the course may be driven by a diverse set of reasons (e.g. program requirements, desire to learn, or considering options). Therefore, there is the potential that student motivation is the driving force for students to succeed within either of the Foundations courses, which is a factor that is independent of student gender.

If the student cohorts do not show any statistically significant differences between them it would be expected that the cohorts follow the same distribution of raw scores; however, where a statistically significant difference occurs, how that is reflected within the raw score distribution should be considered. Typically, due to self-selection it is expected that within physical science courses there will be a greater number of male students who have high ability, but also due to self-selection there is likely to be a greater number of male students who have low ability measures.[157-160,162,163] In comparison, it is expected that female students will not show as much variability within their ability levels, but that they will not have as many students at the highest ability level.[157-160,162,163] This relationship can be seen visually in Figure 17 where the two cohorts being compared are statistically significantly different (as shown within Table 18) (see Appendix 7.11 for the boxplots of all the assessment comparisons).

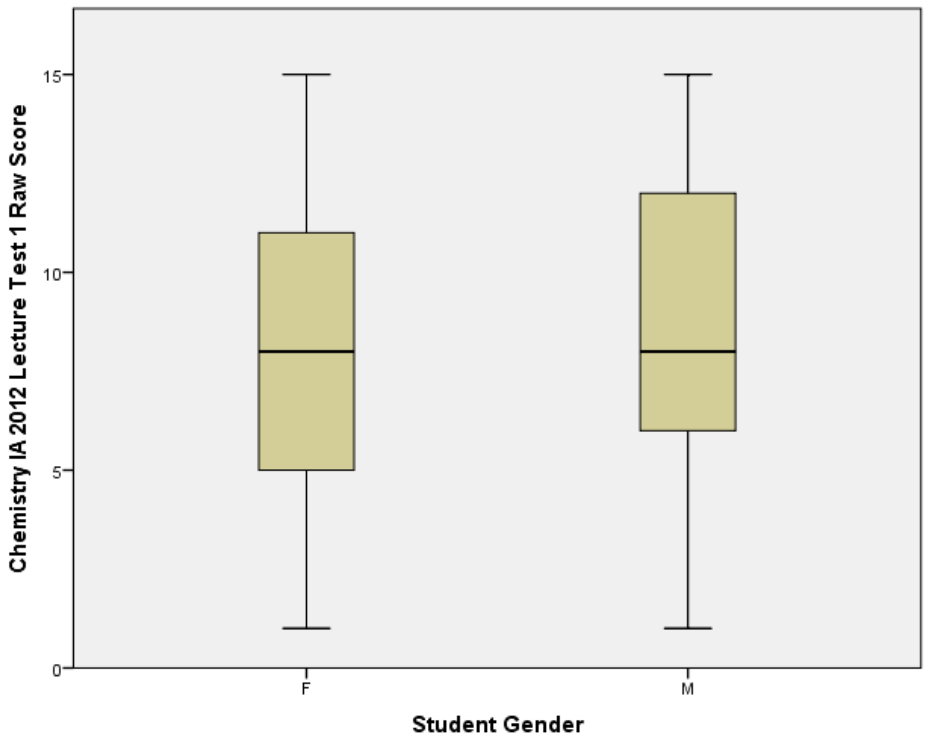

Figure 17: A Comparison of Male and Female Student Cohort Ability based on their Raw Scores within Lecture Test 1 from Chemistry IA 2012

It is not immediately obvious when the two boxplots are compared within the figure how they are classified as statistically significantly different from each other, as the minimum, maximum, and median scores obtained by both student cohorts is the same. The reason that the minimum and maximum scores are the same is due to the limits placed on the scores by the assessment task itself, and within an MCQ assessment containing 15 items it is not unreasonable to expect that within large cohorts (as the male and female student cohorts are) the minimum and the maximum score will be obtained. The same median result for both cohorts is not expected for two statistically significantly

different groups; however, it is important to remember that this is the median result, which means that it is the result obtained by the student in the middle of a sorted list of results. Within an assessment task where the possible outcomes range from 0 -15 it is not unreasonable that the centre result is the same between two cohorts, even if those cohorts are statistically significantly different. Thus, if nothing can be determined from the minimum, maximum, or median of the boxplot, what must be observed is the differences between the quartiles, which informs the distribution of the students across the possible outcomes. The female student cohort quartiles are evenly spread across the entire assessment task, which implies that the results of female students are distributed across all the possible assessment outcomes. The male student cohort quartiles are not evenly distributed, the first quartile and the third quartile are much more elongated than the second and fourth, suggesting that there is a greater range of results within those ranges and more students achieving the same results in the second and fourth quartile. The cluster seen within the second male quartile suggests that a large proportion of the male student cohort obtains results close to the median, and the elongated nature of the third quartile and the cluster of the fourth quartile suggest that the proportion of high achieving male students is greater than the proportion of high achieving female students. Therefore, using the boxplots it can be observed that comparatively more male students are achieving higher results than female students which results in the statistically significant difference determined by the independent sample t-test.

There is the potential that these differences within the student cohort are the result of differences in how each cohort performed on individual items rather than an inherent statistically significant difference between the two cohorts. As these significance tests are based on the raw scores of the students it means that these results are only true when comparing the items using CTT analysis. If the student cohorts are compared using Rasch analysis, the students' ability measures generated through Rasch need to be used instead. Comparing the two cohorts using student ability measures instead of student raw scores can be done using an independent sample t-test. A breakdown of the male and female student cohort differences in all four Chemistry courses can be seen below in Tables 22 - 25.

Table 22: Comparison of Male and Female Student Cohorts undertaking Chemistry IA using the Rasch Ability Measures to determine if Student Ability is Significantly Different between them. Highlighted Cells indicate Observation of a Statistically Significant Difference

Chemistry IA	**2012**		**2013**		**2014**		**2015**	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	467	0.008	446	0.184	471	0.069	502	<<0.001
Lecture Test 2	444	0.093	418	0.012	434	0.010	449	<<0.001
Exam Part 1	506	0.002	503	0.983	505	0.355	544	0.707
Redeemable Exam	485	0.165	486	0.244	496	0.001	523	0.382

Table 23: Comparison of Male and Female Student Cohorts undertaking Chemistry IB using the Rasch Ability Measures to determine if Student Ability is Significantly Different between them. Highlighted Cells indicate Observation of a Statistically Significant Difference

Chemistry IB	**2012**		**2013**		**2014**		**2015**	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	380	0.024	376	0.007	421	0.019	426	0.001
Lecture Test 2	362	0.182	346	0.136	392	0.101	389	<<0.001
Exam	431	0.683	448	0.710	484	0.127	484	0.303

Redeemable Exam	419	0.277	432	0.513	454	0.295	469	0.193

Table 24: Comparison of Male and Female Student Cohorts undertaking Foundations of Chemistry IA using the Rasch Ability Measures to determine if Student Ability is Significantly Different between them. Highlighted Cells indicate Observation of a Statistically Significant Difference

Foundations of Chemistry IA	**2012**		**2013**		**2014**		**2015**	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	257	0.389	307	0.319	250	0.892	292	0.098
Lecture Test 2	265	0.028	253	0.837	221	0.358	234	0.040
Exam Part 1	304	0.439	363	0.133	325	0.690	365	0.756
Redeemable Exam	256	0.632	334	0.344	299	0.691	329	0.116

Table 25: Comparison of Male and Female Student Cohorts undertaking Foundations of Chemistry IB using the Rasch Ability Measures to determine if Student Ability is Significantly Different between them. Highlighted Cells indicate Observation of a Statistically Significant Difference

Foundations of Chemistry IB	**2012**		**2013**		**2014**		**2015**	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	234	0.560	247	0.940	214	0.896	229	0.050
Lecture Test 2	187	0.632	216	0.887	182	0.707	196	0.409
Exam	264	0.103	303	0.292	274	0.465	298	0.026
Redeemable Exam	248	0.985	286	0.685	257	0.994	275	0.792

Analysing the courses with prerequisites it can be observed that Chemistry IA has 7 assessment tasks that display the male student cohort having a statistically significantly higher ability and Chemistry IB has 5 assessment tasks that show the male cohort has a statistically significantly higher ability. Comparing this to the courses that do not have prerequisites it is observed that in both Foundations of Chemistry IA and Foundations of Chemistry IB there were two occasions in which the male student cohort showed a statistically significantly higher ability than the female student cohort. Where the difference in student cohort ability is occurring follows the same expectations that were outlined previously, in that there are more cases within courses with prerequisites likely due to self-selection occurring within the student cohort. An example of two statistically significantly different cohorts can be seen in Figure 18, which compares the male and female student cohorts in Lecture Test 1 from Chemistry IA in 2012 (see Appendix 7.14 for the boxplots of Rasch student ability comparisons for the other assessments analysed).

Similar to the previous boxplot displayed, the minimum and the median are the same for both student cohorts; however, the two notable differences is that the female student cohort contains outliers and the maximum for the male student cohort lies higher than the maximum of the female student cohort. As this boxplot is for the same course and assessment task that was discussed previously within Figure 18 (except using the Rasch student ability instead of the raw scores), many of the observations stated previously remain true. It is the student distribution that needs to be compared between the two boxplots, and based on this it can be seen that while the first and second quartiles of both the male and female student cohorts are very similar the difference occurs within the third and fourth quartile. Within these quartiles the distribution of the male student cohort shows that they have more high ability students than the female student cohort. This can also be observed since the highest ability female students who are considered as outliers within the cohort would lie within the fourth quartile of the male student cohort. It is these differences within

the boxplot that can be used to rationalise why the two cohorts are statistically significantly different from each other.

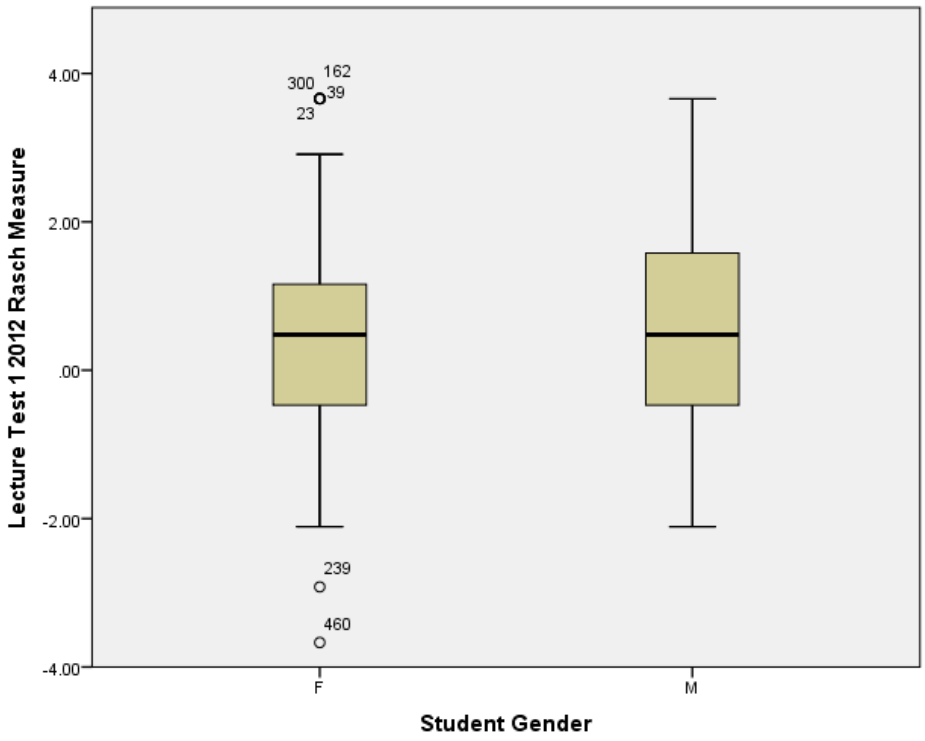

Figure 18: A Comparison of Male and Female Student Cohort Ability based on their Rasch Student Ability Measure within Lecture Test 1 from Chemistry IA 2012

The results of the Rasch comparison of the male and female student cohorts closely match what was seen within the raw score analysis, suggesting that if the goal of an analysis was to simply compare the performance of male and female students there would not be an advantage in undertaking Rasch modelling. If the goal of the analysis is to compare the items, then comparing the results of male and female students needs be done for CTT analysis to determine the expectations that should be had of the items being analysed; however, within Rasch analysis this is not required due to the independence of the student ability measures and the item difficulty measures.

3.4.3 Testing for Gender Differences with Classical Test Theory

There is no methodology described within general CTT analysis to compare the results of the male and female student cohorts on their performance on individual items; however, the assumptions of CTT can be applied to a statistical comparison of the two cohorts on each item. This was done using a chi-squared test to determine if there was a statistically significant deviation between probability of male and female students selecting the correct answer to an item. The null hypothesis within this research was that both male and female students were equally as likely to select the correct option on any item. The effect size (Cohen's *d*) was used to determine how large the difference between the mean of the student raw scores was to gauge the size of the difference between the results of the male and female student cohort on each item. This allows for the items to be evaluated in a non-binary way to provide evidence as to whether any significance that may be observed by the chi-squared test represents an issue that is a threat to assessment validity. These calculations were performed on every item within the assessment tasks being analysed; in addition, the items that were found to show statistically significant differences were gathered to determine how many of them showed statistically significant differences on multiple occasions. It was also considered

whether the items that were identified as showing gender differences were found to be problematic items, as if an item is problematic then it is not a valid item within the assessment and thus needs to be corrected before a judgement can be made as to whether male and female students perform differently on the item. In the same way that the problematic items were considered with multiple years' worth of assessments being analysed, only items that repeatedly showed gender differences were considered as items that needed to be further analysed (except in the case that an item was only asked on one occasion). All the items and their significance levels can be seen in Appendix 7.12, which found 27 unique items out of the 249 unique non-problematic items considered to consistently show differences in how male and female students performed on them.

One of the most obvious outcomes from this analysis is that almost all of the items identified favoured male students, with 18 of the items showing a male cohort bias compared to the 4 items that showed a female cohort bias. Curiously, there were also items that changed which cohort they were biased towards between years, with 5 items identified that favoured either the male or female student cohorts in different years. Reviewing the effect size of these items suggests that in most of the cases the size of the difference between the two cohorts is relatively small despite its statistical significance. Considering the small effect sizes and the evidence that some items change between years it is possible that some of these items are not causing issues within the assessment tasks, as they may only appear due to random variations between the two cohorts. While that is a possibility, it is the consistency with which some of these items appear that is cause for concern, and while none of the items are statistically significant every time they are asked (apart from the item only asked the once) they are still consistent enough that it is unlikely that they repeatedly exhibit gender bias as a result of random variation. This means that these items need to be treated as a threat to the validity of the assessment, as they may be unfairly influencing the results of one of the student cohorts. The issue with items that show gender bias in some way is that it is difficult to use the information obtained from this analysis to improve the items in a way that may remove the bias from them. Therefore, it is important to be aware of which items are causing issues so that attempts can be made to correct the difference observed between the two cohorts, which may require changes to the way that content is taught rather than changes to the items themselves. It may also simply be preferable to remove these items from the assessment task if they are consistent in their bias, as solving the root cause of the issue may require reusing the item and thus risk assessment validity.

There are several considerations arising from this analysis that need to be kept in mind. The first consideration is the difference in student ability that was identified previously. The highest number of items that showed gender differences came from Chemistry IA, which also has the highest number of statistically significant differences between the raw scores of male and female student cohorts within the assessment tasks. This is one of the reasons why items were required to show statistically significant differences on multiple occasions before they were considered for further evaluation, to ensure that shifts in student ability are not the reason for significance. This is a consideration that should be remembered when evaluating each item independently once it has been identified as a potential cause for concern. It is also important to avoid statistical outliers, as it is possible that items may appear statistically significantly different due to random variation within the student cohort; however, it is extremely unlikely that an item will appear statistically significant on multiple occasions due to random variation. Thus, multiple occasions of statistical significance were required before an item was considered to assess male and female students differently. Another consideration needs to be the size of the student cohorts being compared; while they are smaller than the complete cohort (as the cohort is broken roughly in half between male and female

students) the sample size is still large enough that statistically significant differences occur more readily than expected. To compensate for this, the effect size was used as a measure of the size of the significance, and while the two values do share an obvious correlation, the effect size was used as a secondary measure of determining significance. Less of a consideration and more of an observation based on these results is that CTT has identified more items to show statistically significant differences in gender performance than it identified as problematic items. There was very little overlap between the items that were identified to be problematic and the ones that showed statistically significant differences in gender performance: only 1-4 items per course. There is no inherent expectation that there should be more of one than the other, and ideally there are none of either. Thus, it is important that any items that are identified to show consistent statistically significant differences in gender performance, or are identified to be problematic, are addressed and any flaws identified are resolved.

3.4.4 Testing for Gender Differences using Rasch Analysis

Using Rasch analysis means it is possible to determine unique item difficulty measures based on different cohorts of students within the same assessment. By doing this it is possible to compare the item difficulty measure values that are obtained for each student cohort and based on that determine if there is a statistically significant difference in the item difficulty for each cohort. Before breaking down each item individually and determining if any of them are of interest it is possible to visually see the difference in how each cohort performs on an item. This is done using a differential item functioning (DIF) plot, which can either be used to illustrate the item difficulty measures, or to show the size of the difference between two item difficulties. DIF is measured by comparing the item difficulty measures generated by considering only one of the two groups being compared; in this case the item difficulties generated when the male and female student cohorts are considered separately. A difference of 0.50 between two DIF measures is the standard[273,309] for determining if those two measures are statistically significantly different from each other based on Rasch analysis, implying that the two cohorts must be performing significantly differently from each other on that particular item. Figure 19 illustrates the difficulty measure for each item for both the male and female student cohorts, which can be used to determine items of interest before analysing the numerical breakdown of the items. (It should be noted that there is no purpose for items being connected with a line, but it makes the graph easier to follow.) A DIF plot does not have to be used to identify potential items, as the numbers will need to be analysed to accurately determine significance; however, in assessments with a large number of items or to visualise the differences in difficulty, a DIF plot is an effective way to represent the differences.

Based on Figure 19 the items that have the potential for statistically significantly different difficulty measures are item 1 (ΔDIF = 0.55), item 12 (ΔDIF = -0.43), and item 15 (ΔDIF = -0.51). When comparing the items numerically Rasch provides two key pieces of information: the DIF contrast (the difference between the two difficulty measures, ΔDIF), and a *p*-value which within Rasch analysis represents if the observed ΔDIF is due to chance (the null hypothesis is that there is no statistically significant difference between the two DIF values, and therefore it is expected that differences observed are the result of chance deviations between the two cohorts). These values were broken down for every item asked in every assessment task, and the items that showed a statistically significant difference in their ΔDIF on more than one occasion (except for items only asked once) and were not previously identified to be problematic items through Rasch analysis where gathered. The 14 unique items identified out of 178 unique non-problematic items and the values that show difference in the performance of male and female students can be seen in within Appendix 7.15.

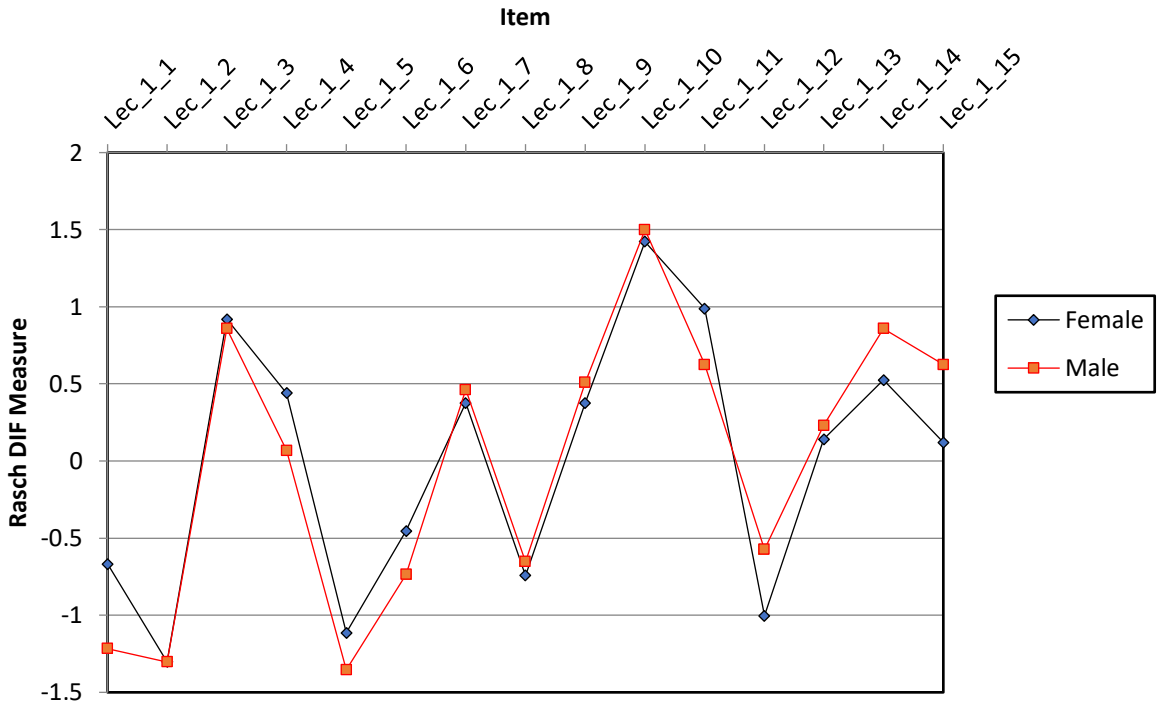

Figure 19: A Line Graph of the Item Difficulty Measures from the Male and Female Student Cohorts in Lecture Test 1 from Chemistry IA 2012

In comparison to what was observed within the CTT analysis the gender biased items are more evenly spread between the two cohorts, with 7 favouring the male student cohort, 6 favouring the female student cohort, and 1 that swaps between the two. The size of the ΔDIF was consistently between 0.45 – 0.60 (items whose ΔDIF rounded to 0.5 were included if they showed a consistent difference), indicating that most of these items are only marginally different enough from each other to be considered outside of the same measure. Even though that is the case the *t*-value on each of these occasions indicates that the difference is not the result of random variation between the two cohorts, and thus some other factor must be causing the differences observed. This means that all the items identified to contain differences between the male and female student cohorts need to be treated as a threat to assessment validity; however, as there are so few of them spread across multiple assessment tasks it is unlikely that they invalidated the assessment tasks being analysed within this research. Similar to the conclusions drawn within the CTT gender biased items, it is difficult to use the information obtained by the analysis to make educated changes to the items that will remove the gender bias. Therefore, these items need to be revised and their performance continued to be monitored, but it is possible that the gender bias is not a direct result of the item itself and thus changing the item may not be addressing the root cause of the issue.

Many of the items that were identified through Rasch analysis to contain differences in gender performance also were identified to be problematic items. As the items were already identified to be flawed, they were not included as gender difference items, which significantly reduced the number of items identified through the analysis. It is possible that an item may be both problematic and show gender differences due to factors that are independent of each other; however, it is best to first solve the underlying issue within the item before attempting any sort of deeper analysis as it is unknown how the changes to the item will impact its performance. It is also very apparent that this list of items identified using Rasch analysis is far shorter than the list of items identified using CTT, as

Rasch identified 14 unique items compared to the 27 identified using CTT. Out of those items 8 of them were found to be statistically significant by both CTT and Rasch analysis, which suggests that those items are likely to be items of particular concern, but it does highlight that there are differences in the results of the analytical techniques. As Rasch analysis does not have to make the same assumptions as CTT (as student ability and item difficulty are independent of each other) it means that there is a higher confidence in the results obtained by Rasch analysis as they cannot be influenced by the student cohort and therefore the results are only reflective of the items themselves.

3.4.5 Deconstructing Gender Differences

Approaching an item that shows gender differences in the students' performance is the same as approaching a problematic item, except instead of looking for inconsistencies in the cohort in answering the item, the inconsistencies between how male and female students are answering the items need to be determined. This is a more difficult process than it is for problematic items, as even though conceptually the process is the same, determining what issues within the item may be causing the differences in student performance is more subjective. For example, when breaking down the stem of an item that shows statistically significant differences between the male and female student cohorts the only approach that can be taken is to evaluate the stem as if it were a problematic item. Not enough is known about the causes of differences in male and female performance to be able to identify anything within a stem that may be advantageous or disadvantageous for one group of students. The stem shown below was found to show gender differences on 2 occasions using CTT analysis and 4 occasions using Rasch analysis out of 8 times that the item was used within assessment. Evaluating the stem does highlight some construction issues with it, as it directly refers to the options drawing the students' focus away from the rest of the stem, but nothing that would be expected to have a significant influence on student performance.

Which one of the following represents the conjugate acid and the conjugate base of the $H_2PO_4^-$ ion?

The stem of this item could easily be reworded as below to remove the reference to the options, giving the students a clear question that can be answered without the desire to check the options immediately.

What is the conjugate acid and conjugate base of the $H_2PO_4^-$ ion?

(A) Conjugate acid: H_3PO_4; conjugate base: HPO_4^{2-}

(B) Conjugate acid: HPO_4^{2-}; conjugate base: H_3PO_4

(C) Conjugate acid: HPO_4^{2-}; conjugate base: PO_4^{3-}

(D) Conjugate acid: H_3PO_4; conjugate base: PO_4^{3-}

(E) Conjugate acid: PO_4^{3-}; conjugate base: H_3PO_4

While this new stem fixes the construction issues seen within the old stem there is no way of knowing if this has addressed the reason for the differences in gender performance. The next step needs to be to look at the distractors of the item to determine if there is a difference in how male and female students are interpreting the options. The distractor analysis for a gender difference item requires slightly more work, as Rasch analysis does not provide an option breakdown for each cohort listed within the assessment. This means that either the analysis needs to be undertaken between male and female students separately or a distractor analysis generated using the raw student results. If Rasch analysis is run separately between the two cohorts then the process is the same as what was described previously within problematic item distractor analysis; however, it is

the difference in male and female students that is of importance. Generating a distractor analysis based on the raw student selection information will provide information listed in Table 26, where not only the counts are included but also the percentage, as the number of male and female students may not be equal and thus using counts alone does not give a reasonable comparison.

Table 26: Male and Female Student Option Selection Rates from Item 8 in Lecture Test 1 within Chemistry IB 2012 where Option A is the Correct Response

Males			**Females**		
Option	**Count**	**%**	**Option**	**Count**	**%**
(A)	136	70.83	**(A)**	95	50.00
(B)	21	10.94	**(B)**	38	20.00
(C)	18	9.38	**(C)**	30	15.79
(D)	17	8.85	**(D)**	20	10.53
(E)	0	0.00	**(E)**	7	3.68

Table 26 shows that in this example male students choose the correct option (A) more often than female students. Based on the evaluation of the differences in the option selection frequency, female students tend to select option (B) more often and have a slightly greater frequency in all other options, with option (C) being the other distractor of note. It cannot be known whether the selection frequencies are a result of male students overperforming or if female students are underperforming; regardless, some attempt needs to be made to bring the two cohorts closer together. Option (B) is the flipped version of the correct response, option (A) (i.e. the acid is labelled as the base and the base is labelled as the acid), which does suggest that perhaps female students are misunderstanding the concepts of what constitutes a conjugate acid and base but understand the relevance of donating and accepting of protons to acid/base chemistry. If this is the case, and it is an issue with conceptual understanding, then there is little that changing the item can do to alleviate this issue, and instead how the course can be changed to address this disparity needs to be considered. Within this item there are not many changes that can be made to the options aside from changing their presentation. Even though it is unlikely that the presentation is causing the issue, without a clearly definable issue present within the options the only reasonable response is to improve every possible facet of the item. Re-evaluating the items after these improvements may show no change to the outcome, in which case the item needs to be reanalysed or simply removed and replaced, or it may solve the issues observed previously. One of these tables should be generated on each occasion that the item shows significance to see if there is a consistent trend in how the student cohorts deviate from each other, and they can also be generated for occasions when the item does not show significance to see if those trends differ on those occasions or simply lie outside of significance. The key to improving assessments and ensuring their validity is continual analysis and evaluation of the results of assessment to confirm that the students and items are performing as they are expected to.

3.4.6 Item Categorisation

The other way in which the items were studied after they had been individually analysed using CTT and Rasch analysis was to categorise them based upon their construction and expectations of the students. Within this research this was done in an attempt to identify trends within item construction that caused items to become problematic within an assessment task, in the hopes that if specific trends could be identified they could help to identify areas within the item that needed to be adjusted. Theoretically, it would also be possible to apply any trends identified to items when

they are being constructed to avoid generating items that have flaws that are known to cause problems within assessment tasks. There was no methodology within the literature that could categorise all the items within an assessment task that would be able to describe their construction and the process that was expected of the students. However, many different aspects of multiple-choice items have been previously discussed and using those different aspects it became possible to generate different categories of interest and factors that could be used to describe how items differ within those categories. Outside of this research, this process has the potential to be used in similar ways, whether it is used before or after the items have been used within an assessment task. It is also possible to utilise only one of the categories if there is something of particular interest that either needs to be included or excluded from the assessment task.

The first part of determining the classification of the assessment is to determine the content that is being covered within the assessment. This is dependent upon the course that the assessment task is used within, as the task may cover only one topic, or it may include several different topics. For example, within Chemistry IA at The University of Adelaide for the assessment tasks analysed in this work, the topics covered are Atoms to Molecules, Energy and Equilibrium, Periodicity and the Main Group, and Transition Metal Chemistry. So, within each assessment an item needs to be assessing one of those topics, which can then be used to evaluate the spread of items to topics. From those counts and using the results discussed within previous sections, Figure 20 can be produced, which uses all the unique items asked in Chemistry IA over the analysis period of four years.

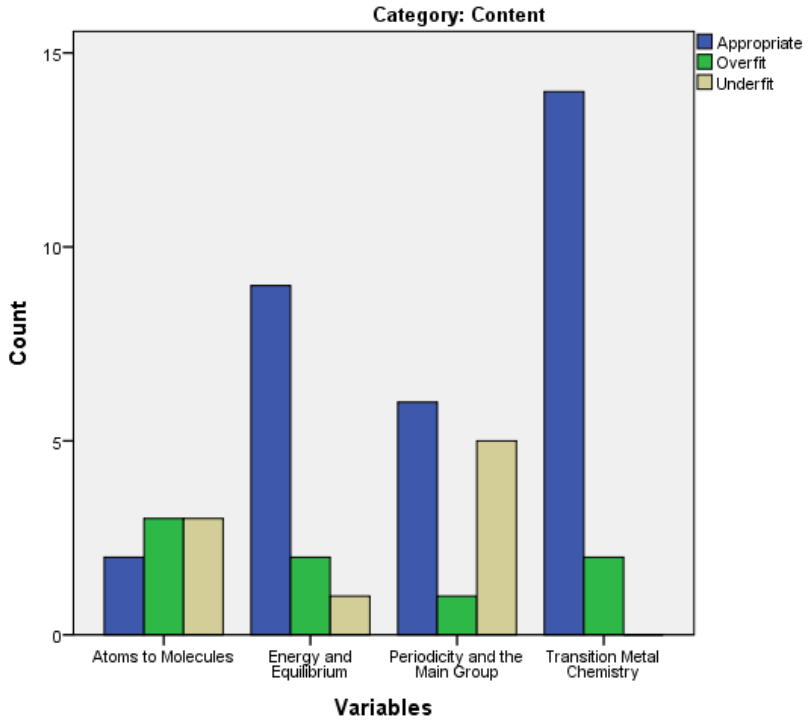

Figure 20: The Number of Items Present for Each Topic Covered Across All of Chemistry IA from 2012-2015

Ideally, each of the items would be evaluated either by the assessors that originally created the items, or with their help to ensure that each item is correctly characterised. It can be seen that in the case of the content covered within Chemistry IA there is a slight concern over the number of misfitting items in Atoms to Molecules compared to the number of non-misfit items. The number of

items identified as underfit within Periodicity and the Main Group is also a potential cause for concern almost matching the number of non-misfit items. While it is not possible in this case to observe how many items of each topic are asked within assessment tasks (as this shows all the unique items used which is not reflective of how many times they appear within the assessments), a possible application of these counts is to ensure that no topic is getting more items than its weighting within the course should allow for (e.g. a topic that is worth 70% of the students final grade should, in most cases, have more representation within an assessment task than a topic that is only worth 10%).

The next consideration for the items is the taxonomy they belong to, which relates to the level of thinking that is required of the students to answer the question being asked. There are a variety of different taxonomies that have been developed for the purpose of student evaluation and assessment; however, many of these taxonomies relate to the students specifically and cannot be used to evaluate the construction of items. The two taxonomies that were used in this work were Bloom's Revised Taxonomy[71] (which for the purposes of MCQ item evaluation, matches closely with Bloom's Taxonomy) and the Structure of Observed Learning Outcomes (SOLO)[72], both of which are briefly outlined below.

- Bloom's Revised Taxonomy (Cognitive Domain)
    - Remember
    - Understand
    - Apply
    - Analyse
    - Evaluate
    - Create
- Bloom's Revised Taxonomy (Knowledge Dimension)
    - Factual
    - Conceptual
    - Procedural
    - Metacognitive
- Structure of Observed Learning Outcomes (SOLO)
    - Prestructural
    - Unistructural
    - Multistructural
    - Relational
    - Extended Abstract

Bloom's revised taxonomy uses two dimensions to classify the level of thinking required to complete a given task: the cognitive dimension represents the complexity of the task and the knowledge dimension broadly represents how abstract the information within the task is.[71] The complexity of a task can also be considered as the order of thinking that is required to complete it, and thus applying the cognitive domain to assessment items means that it is representative of the level of thinking required from the students to answer it correctly. Lower order thinking represents the ability to remember facts and explain what those facts inform. Middle order thinking is being able to take facts and apply a known process to determine potential outcomes or being able to deconstruct a piece of information into its base components and separate them based upon their merit. Higher order thinking is being able to judge all the pieces of information presented and organise them based on their relative importance and accuracy before using that information to rationalise the

potential outcomes of what was observed or should be expected. Applying the knowledge dimension to assessment items categorises how abstract the information required to answer the item is. Factual requires knowing terminology and specific details, conceptual requires knowledge of classifications, theories, and principles, procedural requires knowledge of skills, methodologies, techniques, and the requirements of all of them, and metacognitive is the requirement of knowing how much is known by the individual. Not all these classifications are applicable within all assessment tasks, and as MCQ assessments were being analysed within this research, it meant that the highest orders of thinking (evaluate and create) were not assessed.

The Structure of Observed Learning Outcomes (SOLO) is a different learning taxonomy that was also used to categorise the items due to its common application within education and the potential to apply it to assessment items.[72] The SOLO taxonomy describes increasing levels of complexity in the students' understanding of topics using 5 different stages of understanding. Prestructural means that the student does not understand the core concepts, unistructural means the student focuses on only one relevant aspect, multistructural is when the students are able to discuss several relevant aspects but are unable to link them. Relational is when the students are able to integrate different aspects together and thus can describe the aspects and confidently link them together, and extended abstract is being able to take the integrated aspects and apply them in new ways or to a different topic. This makes SOLO a student-focused taxonomy, and thus when applying it to assessment items it needs to be looked at in a different way. Within this research an item was placed within the different SOLO categories based upon the level of student understanding that would be required of them to answer the item. Therefore, if an item required the student to apply a specific equation, this would be considered unistructural, as all the student needed was the knowledge of the equation; however, if the student needed to apply that equation and then use that result to explain why a particular outcome was observed this may be a relational item. The level of understanding required from the students typically correlates to the order of thinking that is required from the item, thus a relational item likely requires higher order thinking; however, this is not always true and needs to be considered on an item by item basis. Like Bloom's revised taxonomy, not all the categories within SOLO are applicable to MCQ assessments, and thus it was not possible to observe any extended abstract items.

Each item was assigned to one of the categories from each of the three different taxonomies and using the counts for each of the taxonomies, a description of the level of thinking that the assessment requires of the students can be deduced. An example of how this process was undertaken is shown using the two items below.

(1) *Sodium emits light of wavelength 690 nm when heated in a flame. What is the frequency of this light?*

(A) $2.07 \times 10^{11}\ s^{-1}$
(B) $4.35 \times 10^{5}\ s^{-1}$
(C) $4.35 \times 10^{14}\ s^{-1}$
(D) $207\ s^{-1}$
(E) $1.04 \times 10^{14}\ s^{-1}$

(2) *Consider the following chemical reaction*

$$N_2(g) + 3H_2(g) \rightarrow 2NH_3(g)$$

The reaction indicated above is thermodynamically spontaneous at 298 K, but becomes non-spontaneous at higher temperatures. Which of the following is true at 298 K?

(A) ΔG, ΔH and ΔS are all positive
(B) ΔG, ΔH and ΔS are all negative
(C) ΔG and ΔH are negative, but ΔS is positive
(D) ΔG and ΔS are negative, but ΔH is positive
(E) ΔG and ΔH are positive, but ΔS is negative

Using the definitions of the taxonomies previously discussed and upon reviewing item (1) it is reasonable to define this as an understand, conceptual, and unistructural style of item. These categories were decided as the students are required to link the information given within the stem to the equation that describes the relationship between frequency and wavelength, which means there is only one concept involved in answering the item and involves student understanding of a conceptual relationship. Item (2) could be described as an analyse, procedural, and multistructural item as the students are required to utilise multiple concepts independently to determine how an outcome is possible based on a procedural process.

Figure 21 shows the breakdown of all of the unique items used within Chemistry IA, and hence can be used to gain an understanding of what level is required of the students throughout that course and what levels of thinking are causing the most issues when they are required by the students to answer the item. (The 'evaluate' and 'create' from Bloom's Revised Taxonomy, and 'extended abstract' from SOLO are omitted from Figure 21 as they were not seen or expected within the items).

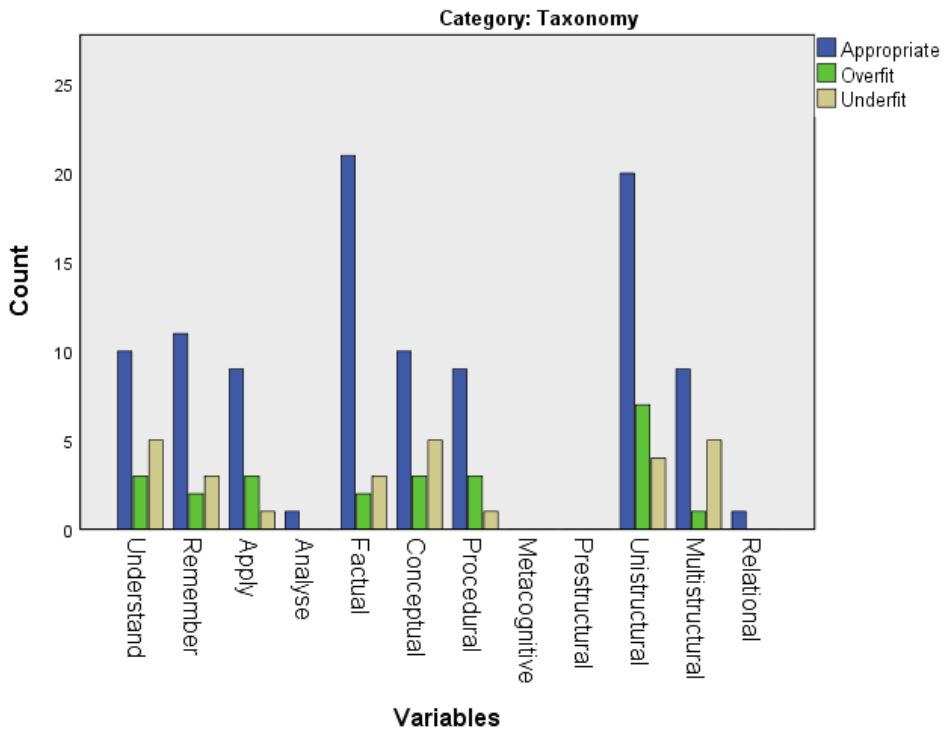

Figure 21: The Number of Items Present for Each Taxonomy Based on all of the Unique Items used in Chemistry IA from 2012-2015

No single area is responsible for most of the problematic items, and there is a reasonable spread of items over different orders of thinking, as shown by the distribution of the items across the cognitive domain. The items seem to trend towards a lower level of understanding being required by the students, which is represented by the fact that more items lie towards the start of the SOLO taxonomy categories. The shape that this graph is expected to take is dependent on the type of assessment task that the assessors are attempting to construct. If the assessment is constructed with the intention of using MCQs to assess students on their knowledge of the facts and concepts relevant to the topic then it should be expected that the results would skew towards the lower order thinking and lower levels of understanding. If the intention of the assessment is for students to apply those concepts and analyse information from the topic then the results should show the items favouring higher order thinking categories and require a higher level of understanding. In Chemistry IA it is expected that the items assess a broad range of thinking categories, and thus it is no surprise to see a spread of different levels of thinking and understanding being assessed within the graph in terms of both Bloom's Revised and SOLO taxonomies. Depending upon the purpose of the assessment task analysis, one of the taxonomies may be more relevant, and ideally Bloom's revised taxonomy should be viewed in two dimensions to fully capitalise on its functionality. Within this research more was learnt from Bloom's revised taxonomy than SOLO as it was more applicable to MCQ assessments; however, there is an overlap observed between the two taxonomies as they are both different ways of representing the level of understanding that is required to answer the item.

All the categories discussed other than the topics and taxonomies categories have the factors being considered generated by rationalising what is important to assessment tasks and what may have an influence on their outcomes, and thus this allows for adaptations to be made as required. This is because not all factors are relevant to every assessment task, and therefore there may be factors that are not included that are crucial to an assessment task, or there may be factors crucial to this analysis that are not relevant to the work of others. The reason for this is that these factors were generated specifically using chemistry assessments, and thus factors that were important to chemistry were focused on. However, many of these factors are specifically related to MCQs in general and thus ideally this system can be easily adapted to courses outside of chemistry through careful consideration or iteration of the system to focus on factors that are influential to the course being analysed.

Once the order of thinking that is required from an item is known, the next step in this process is to consider what skills the item requires from the students for them to be able to determine the correct option. The consideration of what is required of the students means this category includes influences based on the basic skills required, the knowledge of the students, and the construction of the item. The item type is categorised based on several subcategories that each explain a different aspect about the item and how it is constructed. The first subcategory that was generated was the level of thinking that the item requires to be answered, and thus the categories within this group closely resemble ideas presented within Bloom's revised taxonomy.[71] The next subcategory represents the type of thinking that is required by the students when they approach the item, and thus this group closely represents the different styles of thinking that individuals may utilise as the items usually lend themselves to one more than the others.[334] The next consideration within this category is if there is a methodology that is encouraged by the item in how the students should approach determining the correct answer. The last subcategory represents how familiar the students are expected to be with the item and/or what process that needs to be followed to determine the correct answer and thus ranges from items they consistently practice with to completely new items. Not all the subcategories are applicable to every item, and an item may

match several factors included within a subcategory (except familiarity). All the factors and their subcategories that were considered for item type were:

Level of Thinking:

- Recall/Recognition (simply being able to answer the item from knowledge of the topic)
- Specific Knowledge (requirement for the students to know information not provided within the item that is critical to answering the item but not the answer to the item)

Type of Thinking:

- Comprehension (if key information needs to be extracted from the item stem to answer the question being asked)
- Visual (an important visual aspect is within the item in some way, either provided or generated by the student)
- Quantitative (requires the use or manipulation of numerical information)

Item Approach:

- Application (the specific use of concepts or equations to answer an item)
- Analysis (break down and explore ideas given with the item)
- Logical Reasoning (the students making connections based on knowledge)

Student Familiarity:

- Textbook Item (a practiced item that the students have seen before)
- Novelised Textbook Item (a familiar process that the students are required to follow with a different context or a new approach)
- Novel Item (unlikely that the students are familiar with the item or the process they are required to undertake to obtain the answer)

Each individual item will belong to multiple different categories within this analysis, and the same examples (page 127-128) can be used to show how these categories are determined. Item (1) requires the specific knowledge of the equation used to calculate frequency (if the equation is provided to the students this category does not apply), it is a quantitative item as it requires the students to apply numerical information. It is an application style item as the students need to utilise the equation to determine the answer, and it is a textbook style of item as the students would be familiar with the process they need to apply. Item (2) also requires specific knowledge of Gibbs free energy and the equation used to calculate it (also may not apply if the equation is provided), it requires both comprehension and quantitative types of thinking as the students need to be able to understand the different circumstances discussed within the stem and applying it requires calculating how each thermodynamic property influences Gibbs free energy. Approaching this item requires the students to either use analysis to determine the answer based on the what option can result in the outcome described or logical reasoning to determine how the outcome is possible without using the item options. This is a novel item as the students have likely never been asked to interpret Gibbs free energy in this way previously.

A graph of the items used within the assessment tasks can be generated to view which type of items are the most utilised. Figure 22 shows the type of items that were used within Chemistry IA assessments at The University of Adelaide from 2012-2015 and shows a clear spread across all the item types that were considered.

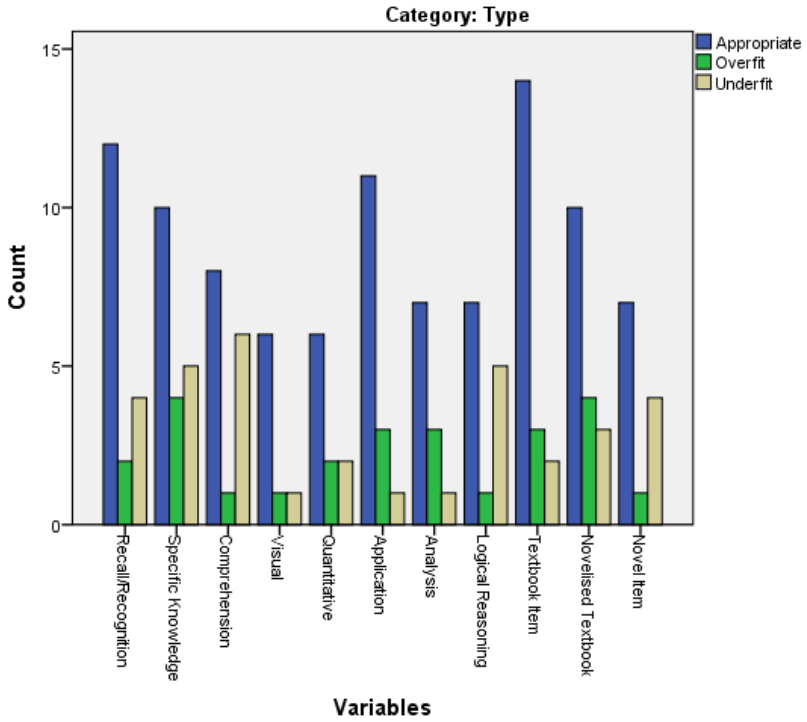

Figure 22: The Number of Times Each Item Type was Present within an Item Using Unique Items from Chemistry IA MCQ Assessments from 2012-2015

Analysing the figure shows that the 'comprehension', 'logical reasoning', and 'novel item' item types were the only three categories that showed several problematic items that closely matched the number of appropriate items. It should not be a surprise that 'logical reasoning' and 'novel items' are two of the categories most associated with underfitting items (items that are less predictable than expected, typically caused by higher achieving students having difficulty or low achieving students having unexpected success), as these categories tend to require the students to take an approach that they are unfamiliar with or may require them to comprehend information differently than their previous experiences. The reason that the number of 'comprehension' type items have a relatively high number of underfitting items is not as clear, but it is possible that the underfit may be caused by 'comprehension' items that require the students to read large amounts of information that may confuse them. To determine if that is true, the problematic items themselves need to be analysed to determine what is causing the issues, as it may simply be coincidental that problematic items fit within these types of items. What is expected of the assessment task also needs to be considered when viewing the results of this data, as the purpose of the assessment task should inform to some extent what item type is planned to be used within the assessment. Within Chemistry IA it was observed that there is a large spread of item types; however it needs to be remembered that as this shows all of the unique items used over the period of four years, it is possible that the results seen in Figure 22 are not reflective of the individual assessment tasks.

The next step is to consider how the item is presented to the students and what information is given to them when they read the item. The item presentation considers the information given to the students within the stem, in what form that information is given, and if the information is being provided within the stem or the options of the item. The factors considered are informed by the content that is presented to the students, the item construction, and how the item informs student

approaches. The categories included here are generated based upon repeated application of this process to determine if all the items are adequately described by the factors listed. Thus, this category is the most likely to change between different assessment tasks or if it is applied to different courses as some of the factors considered are very specific to the content being assessed. For example, as the MCQ assessments being analysed are from a chemistry course factors such as chemical structure and chemical formula need to be included as seen within some of the items present. However, it is unlikely that those factors will be utilised if this process is applied to an assessment task for a music course and instead other factors more relevant to the topics being assessed within that assessment will need to be considered and used.

The factors generated can be broadly placed into two subcategories: presentation of item information and item construction. The presentation of item information is the most susceptible to changes between assessment tasks and was generated based on common ways that important information is imparted to the students within the item. The item construction factors are less likely to change as they are related to important consideration in how an item is worded and presented to the students. These factors are based on important item writing guidelines,[45] almost all of which are specifically related to the construction of MCQ items and thus if this process were to be used for a different assessment format many of these factors would need to be revised. The list of factors considered for the presentation of an item is:

Presentation of Item Information:

- Mathematical Equation (a mathematical equation is given within the item for use)
- Mathematical Information (specific numbers are given with the intention of them being used in any way by the students)
- Chemical Formula (the inclusion of a relevant chemical formula within the item)
- Chemical Structure (the structure of any molecule being shown in any form)
- Chemical Equation (a balanced reaction between molecules)
- Chemical Reaction (a chemical reaction with the structures presented)

Item Construction:

- Direction [Stem or Options] (where are the students told what they need to obtain?)
- Context [Stem or Options] (how is the item relevant to the assessment?, i.e. is the section of content being assessed clear from the stem or the options presented)
- Distinct/Vague Question (is the item answerable from the stem alone?)
- Stem Length [Short or Long] (does the stem contain an excess of information?)
- Option Referral (if the stem directly mentions the options presented)
- Inherent Two-Step Item (two steps required due to how the item is constructed)
- “Complete-the-Sentence” (if the options lead directly on from the stem into one sentence)
- Supplementary Resources (if the item provides students with additional resources to use)
- Item Hint (any indication of how the students may need to answer the item)

Applying the same two item examples as previously described (page 127-128) item (1) presents the students with mathematical information within the stem and no other forms of information are provided. The construction of item (1) means that the direction and context is provided within the stem, which asks a distinct question using a short stem length. Item (2) presents the students with a chemical equation, and even though there are measures of temperature provided within the stem this is not considered mathematical information as it is not expected that the students will use those

numbers in any way. The context and direction for this item is given within the options, as without the options the students would not know what that information is expected to be used for, and hence the item is also a vague question that has option referral within the stem. The stem itself is considered to be long compared to other items, as it provides two distinctly different pieces of information in multiple sentences.

This list of categories will likely need to be changed to fit the topic that an assessment task is covering, as the first six factors listed may not be present within assessments for some topics. These should be replaced with other factors that are continuously included within their assessments. The breakdown for how the unique items in Chemistry IA were presented can be seen in Figure 23, where there is a clear preference for some presentation methods over others.

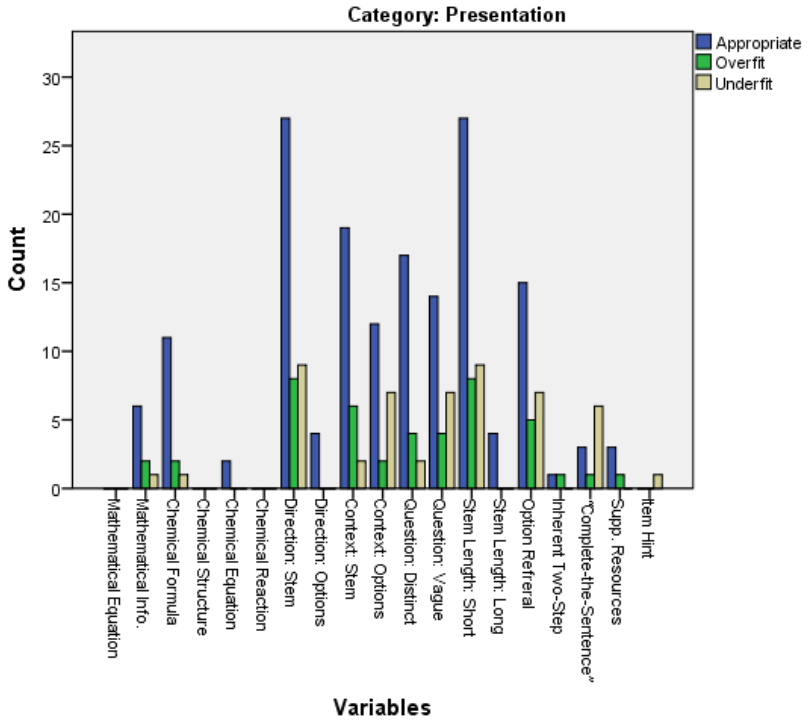

Figure 23: The Number of Times that Items were Presented to the Students Analysing all Unique Items from Chemistry IA MCQ Assessments from 2012-2015

The purpose of the assessment task should inform the expectations of the item breakdown; however, in the case of item presentation the results are not necessarily reflective of how an item functions but rather what is used to inform the students of the item's requirements. For example, the inclusion of a chemical formula does not give information about what the item is asking of the students, nor does it give any information about how that relates to the item itself. What it can be used to inform is if the students are repeatedly being presented with the same kind of information, which may or may not match the expectations of the assessment task. The other information that can be gained through these categories is if items are continually presented with structural flaws and the impact that those flaws have on the assessment. Categories with both stem and options are expected be presented within the stem, as this follows item construction guidelines for MCQs. This does not mean the item cannot be valid if that information is presented within the options, but it does mean more care should be taken in how that information is presented, as whenever the information is presented within the options there is an increase in the percentage of items in that

category that appear problematic (observed by comparing the categories seen within the figure). Similarly, the inclusion of extra factors in an item change how students interpret the item ('complete the sentence', two-step item, hints, supplementary information, option referral), as it changes the information that is available to them and the strategies that they can attempt to employ, and as a result causes large amount of problematic items compared to the items that appropriately fit, as was seen within the data analysis. The categories of presentation should be viewed to be a way of determining if the content being presented to the students is appropriate for the assessment, and if the item construction continually moves away from item writing guidelines.

All the categories discussed above relate to the item and the way that it is constructed, but the approaches that the students can take to determine the correct option need to be considered. While it is subjective to consider the approaches that the students may take to obtain the answer it is still an important consideration. While there is likely an intended process for the students to follow (e.g. apply an equation, use a concept, manipulate information, etc.) the other processes that the students may take need to be considered as well. These categories can be used to determine if the assessment is asking the students to undertake different processes, or if the items are relying on students repeatedly undertaking the same process in different contexts. What this tells the assessors is dependent upon the expectations that they have of the assessment, as it is possible that assessors want the MCQ assessments to be used purely for recall and concept application and other types of assessment are intended to be used to assess other processes. The categories were viewed as the broad approaches that the students could take to determine the correct answer to the item to ensure that all the possible approaches were included. The factors can be considered in several different subcategories of different forms of approaches that the students could take to answer an item. The first subcategory is if the students simply generate the answer based on either their own knowledge or some form of process that can be applied to generate the correct answer independently of the options presented to them. Alternatively, the item or the students may make use of the presence of the options and compare and evaluate the options to determine what they believe to be the correct option. This subcategory needs to be carefully considered as this is bordering on a student answering methodology; however, some items specifically require the students to evaluate the options presented to them to determine which one of them best answers the question presented to them. Within this research there are two factors represented within this subcategory, option comparison and option evaluation, both of which are slightly different; however, both of them may represent the requirements of an item or the students using an answer strategy and thus this subcategory needs to be considered carefully as do the items that fall within it. Another subcategory is if students are required to or choose to apply a more critical style of thinking to answer the item, either through generating new information based on what is presented that can be used to answer the question or simply through rationalising what they believe the correct answer should be through their own knowledge and logic. These types of processes tend to be difficult within MCQ assessments as there is no way for students to demonstrate how they have gone about their reasoning, and thus they only ever receive marks if they are correct. Therefore, any items within these categories should be considered for their potential to be constructed as a longer form item where the students can clearly display their thinking. The last subcategory considered is if the students are required to utilise resources separate to the assessment task itself, which usually refers to something akin to a formula sheet but depending on the assessment task it may be some other form of resources given to the students to use. Many items will have many different factors that can be used to describe the processes that the student may employ, as different students will have different ways in which they approach the assessment task and the items within it. The processes that were considered were:

Answer Generation:

- Recall/Recognition (knowledge of the correct answer to what is being asked)
- Knowledge Application (using knowledge to determine the answer)
- Concept Application (applying a concept to a circumstance to obtain an answer)
- Mathematical Calculations (general basic calculations)
- Equation Formulation (either the selection or generation of a relevant mathematical equation)
- Equation Evaluation (using a mathematical equation to rationalise outcomes non-numerically)
- Equation Application (the use of a mathematical equation to obtain the answer)

MCQ Answer Strategies:

- Option Comparison (comparing each option to the others for what answers the item best)
- Option Evaluation (individually appraising each option based on the information within the stem and the student's knowledge of the topic)

Critical Thinking:

- Deductions/Reasoning (based on the information given, logically be able to connect the ideas presented with personal knowledge to answer the item)
- Information Generation (completing a task that does not answer the item but is required to find the correct answer [e.g. chemical equation, minor calculation, etc.])

Additional Resources:

- Reference Usage (use of provided reference is recommended/required)

Using the same examples (page 127-128) again item (1) can be answered only through the use of equation application and hence none of the other factors are relevant unless the equation is provided within a separate formula sheet (as it was within this assessment task) provided to the students which means that reference usage would also be relevant to the item. Item (2) requires the students to undertake equation evaluation to determine how an outcome is possible, and it may involve option evaluation if students attempt to determine which option provides the scenario that can achieve the desired outcome. It is also possible that students use their reasoning to determine how the outcome is possible based on the factors that can be manipulated within the equation, and it may involve reference usage if the equation of Gibbs free energy is provided within a formula sheet (as it was within this assessment task).

The processes that students are expected to apply may change with the topic or course being assessed, as it may not require any mathematics and thus the mathematical processes described above could be replaced with something more appropriate for that assessment. The breakdown for all the unique items used within Chemistry IA at The University of Adelaide can be seen in Figure 24.

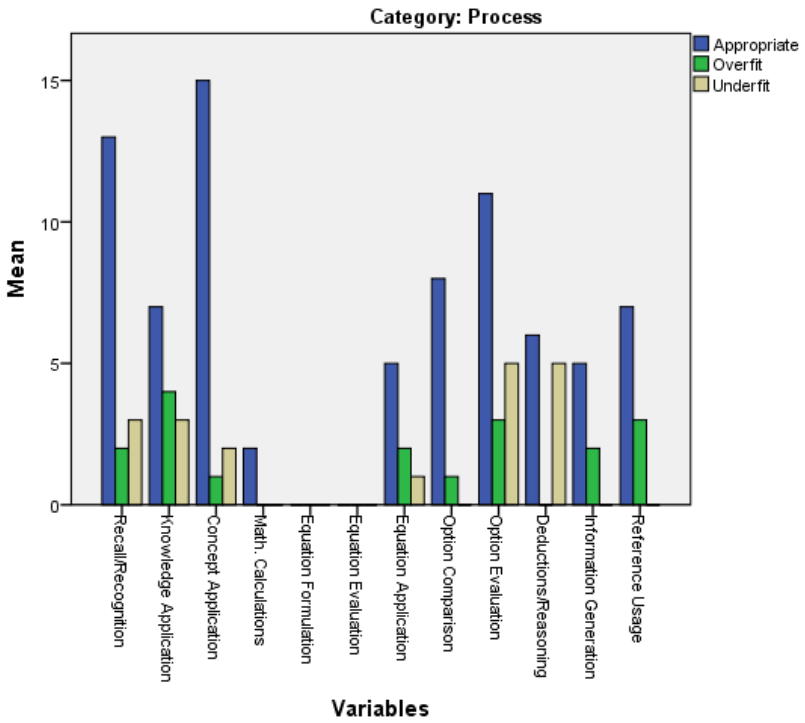

Figure 24: The Number of Times that Different Processes Could be Applied by the Students in all of the Unique Items from Chemistry IA MCQ Assessments from 2012-2015

While the results are not reflective of each individual assessment used, the spread observed should be expected of assessment tasks that attempt to assess the students in a broad range of topics and in different orders of thinking, as the items are distributed throughout almost all of the categories. Both 'option evaluation' and 'deductions/reasoning' stand out as processes of concern due to the high number of problematic items that fall within those categories, and it is likely that the same items may be present within both of those categories. If one of the processes that students are expected to undertake, or have the potential to undertake, is 'option evaluation' it signals a potential issue with item construction, as it means that students are not able to answer the item through the stem alone, thus the item is not adhering to MCQ item construction guidelines and may result in student behaviour that undermines the purpose of the assessment task. 'Deductive/reasoning' style items are expected to be more difficult items within an MCQ assessment as they require higher order thinking and a deeper level of understanding and hence assess the students on higher levels within each taxonomy. One of the issues with having these items within the MCQ assessment tasks analysed within this research is that students cannot obtain marks for showing their process, and thus it may be unclear if the students are struggling with these items or if the students are simply reaching a different conclusion while believing they are on the right track.

While the item difficulty can be evaluated once the assessment takes place, and experienced assessors will be able to give an estimation of how difficult they expect the students to find the item, determining the complexity of the item can give information about the cognitive load that the items are placing on the students. Determining what makes an item complex involves considering what can change between items within an assessment, and what factors will increase the cognitive load on the students when answering the item. For example, the number of options presented to the students is not a good indicator of item complexity, as even though it is more information that the students need to process, the number of options should not change between the items, making it a

consistent amount of cognitive load throughout the assessment. Instead, two subcategories were used to describe the potential for the items to place extra strain on the students. The first potential strain on the students that may change between the items is the number of factors that the student needs to consider within the item itself. These factors are an attempt to describe the difference in cognitive load between two items, and thus consider the three cognitive process that are most likely to change within an MCQ assessment and are thought to place additional cognitive load on the students (approaches, steps, and concepts).[216] The other subcategory relates specifically to MCQ assessments and the differences that may be present between two MCQ items that require the students to process more information. Thus, these factors of answerability and multi-mark relate to the construction of MCQ items and how changes in the construction may require the student to undertake additional tasks that increase cognitive load.[45,216] The factors that were considered are:

Student Cognitive Load:

- Number of Approaches (how the students can answer the item)
- Number of Steps (the number of significant tasks required for the students to answer)
- Number of Concepts (the number of theoretical considerations of the student when answering)

MCQ Construction:

- Answerable from Stem (if an answer can be generated without looking at the options)
- Multi-Mark Potential (if more than one mark or partial marks could be assigned to the item)

In this research, instead of determining the exact number of approaches, steps, and concepts, the options of 'one' or 'multiple' were used. This was done both to simplify the categorisation of individual items, and because the exact number can change depending on the student's own individual methodology, and thus it is likely that the numbers will vary between students. There is potential to add to this depending on the construction of assessment as factors such as the inclusion of supplementary information and figures that need interpretation could be important considerations of item complexity, but they are not utilised frequently enough to justify their inclusion within this research. Using the same examples (page 127-128) item (1) only involves one approach, step, and concept as it only requires the students to undertake a single calculation, and it is answerable from the stem alone. Item (2) cannot be answered from the stem alone, and is not considered to have multi-mark potential as there is no information that the students are required to generate that is not reflected within the answer options. Item (2) requires at least two approaches (option evaluation or logical reasoning), involves multiple steps with either approach (evaluating each option or applying logical reasoning to every factor within the equation), and requires the students to have knowledge of concepts relating to each of the different factors involved within Gibbs free energy. This is an example of the subjective nature of these categorisations, as it could be argued that this item only requires knowledge of Gibbs free energy and hence this should be considered as only one concept. However, within this research an item that only contained one concept needed to represent a simple concept that did not involve other factors (such as what is presented within item (1)), and therefore as item (2) does require knowledge of the factors that influence Gibbs free energy it was considered to contain multiple concepts. This is an example of how the methodology may be applied differently within different assessment tasks and highlights why it is important to clearly define what each category represents before it is applied.

Figure 25 shows the complexity of all the unique items used within Chemistry IA at The University of Adelaide and can be used to evaluate if cognitive load is a concern within the items.

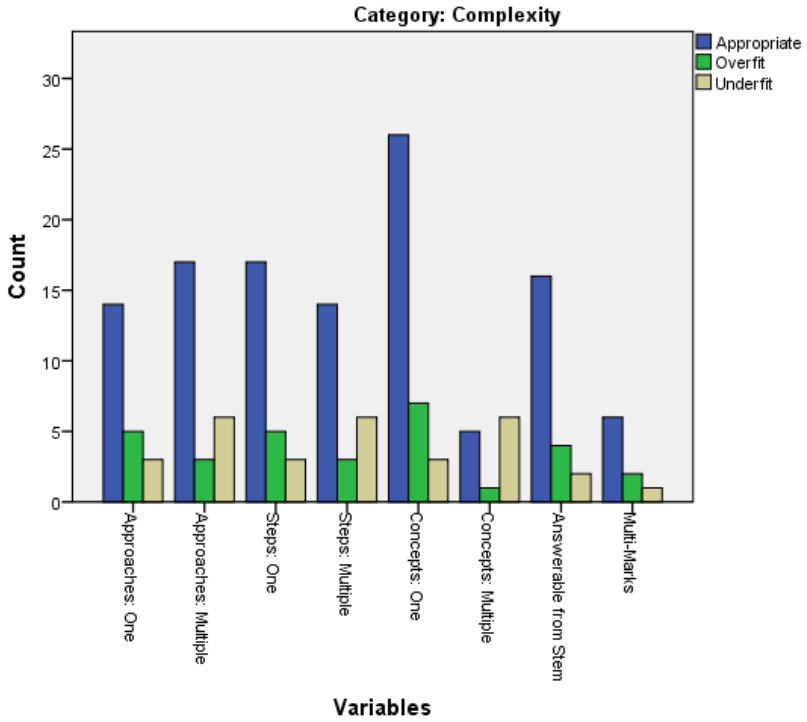

Figure 25: The Number of Times that the Items showed that level of Complexity in all of the Unique Items from Chemistry IA MCQ Assessments from 2012-2015

As Figure 25 is representative of all the items used and not individual assessment tasks, it cannot be used to inform assessors of the cognitive load within any one assessment task, but it does give an overview of the trends of what factors can cause problems within an assessment when considered alongside the items that have been previously identified as being problematic within the assessment task. The observed spread of the data should be expected of assessment tasks that assess the students on a wide range of different items at different levels (as has been observed for this set of assessment tasks within all the categories discussed previously), and several of these categories are closely linked to each other (i.e. commonly items with 'one approach' will only require 'one step' as these items can usually only be approached in a linear manner that does not allow the students to attempt different methodologies [e.g. use an equation, complete a chemical reaction, recalling facts/information, etc.]). The only notable consideration that can be seen from the figure is that there are more underfit items associated with 'multiple concepts' than there are non-problematic items associated with it. Items that contain 'multiple concepts' means that the students are likely to be required to take multiple deliberate steps with no marks given for partial workings. Having 'multiple concepts' within an item also means that there are more considerations for the students that they may either misunderstand or make a mistake when applying them. Thus, it could be theorised that there is the potential that students of higher ability levels are more likely to make a mistake on the multiple concept items than any other type of item resulting in underfit items; however, there is no way to be sure from this analysis if multiple concepts is the cause of underfit or if it is some other unrelated factor. The other factors within the figure of 'answerable from stem' and 'multi-marks' are used to determine the style of items being used within the assessment consistently, neither of which seemed to be related to any issues within this analysis. 'Answerable from stem' items mean that the item gives the students a clear question to be answered, which should be expected of most MCQ items as it is a commonly recommended construction guideline

and does not seem like it influences the items to be problematic based on this analysis. 'Multi-mark' items mean that the item could be broken down into several parts that are each worth their own mark (or at have partial credit), and similarly these items are not seen to have any inherent issues based on this analysis, but they should also be considered if they are appropriate for the assessment.

The only other aspect to consider when evaluating items is whether there are any obvious issues within the item that can be identified by evaluating its construction. This category needs to be carefully considered, as while all the categories discussed can be somewhat subjective, this category is the most subjective. This is because it is always easier to find issues within the item construction that can be exploited by the students when all the factors surrounding that item are known, particular the answer option. Connections that the students will not be able to make within an assessment setting can be made when trying to identify issues due to the additional information known, and thus items may be tagged as potentially problematic when there are no issues with the item construction. All of the issues considered are based upon common concerns within assessment tasks,[45] some of which are more related to MCQ assessments, but all of them should be considered within any assessment task as there is the potential for them to be exploited by the students. The potential issues considered were:

- Cueing Potential (if the answer is indicated in any way within the item)
- "Gameable" Items (potential for 'test wiseness' to impact result)
- Unrelated Assessing (if the item has aspects that are not related to the item)
- Potential Item Construction Issues (any other issue that may be associated with an item)

Ideally items are reviewed before assessment tasks are utilised and thus any glaring issues within the assessment items can be addressed before the items are presented to the students. However, simply because there is potential for an item construction flaw does not mean that the item will not perform as it is expected to within the assessment. Reviewing the examples (page 127-128) item (1) does not have any flaws as it only provides enough information to ask and question and provide the information required to answer it. Item (2) does contain potential flaws, as it requires the students to evaluate the options which may result in 'test-wiseness' influencing student outcomes; however, as the number of ways the equation can be rearranged is limited, the students will not gain any additional information from evaluating the options compared to attempting to determine how the outcome is possible based on the equation alone. It can be seen in Figure 26 that not all of the items that were thought to contain potential flaws were identified as problematic based on item analysis done within the previous section, which means that the flaws do not have an influence on the students, the students were unable to exploit those issues, or the identified flaws do not cause any issues within item construction.

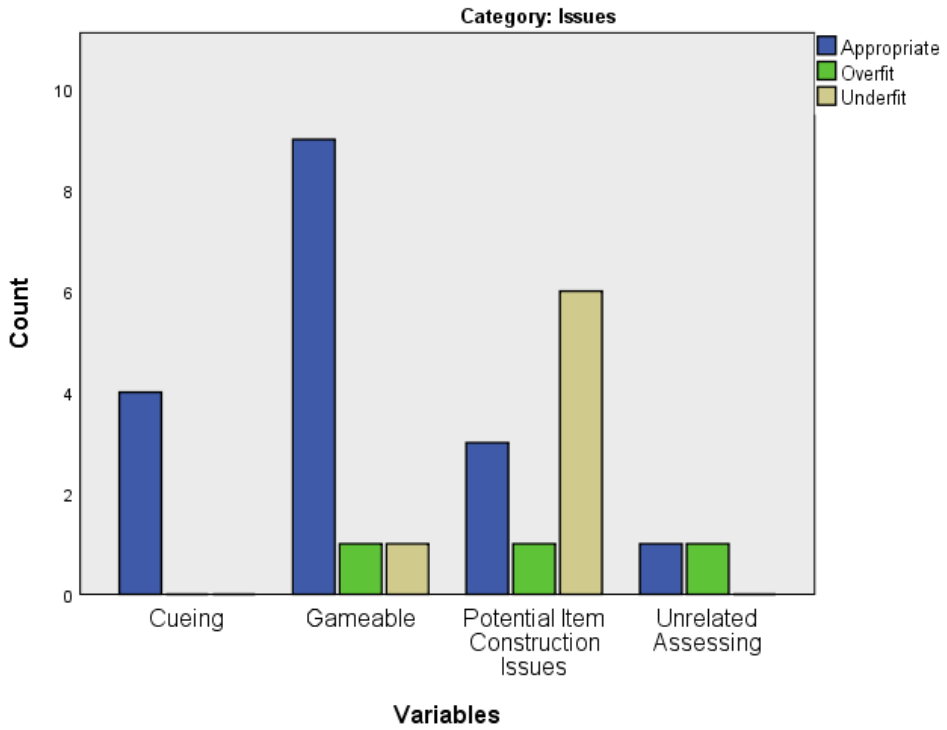

Figure 26: The Number of Times that there were Potential Item Flaws within the Unique Items from Chemistry IA MCQ Assessments from 2012-2015

Reviewing the figure it can be seen that this is true for all of the categories except for 'potential item construction issues', which is the broadest category as any flaws that are thought to be a concern that do not belong within the other categories are placed there. Typically, the flaws seen here relate to issues with how the stem or the options are presented to the students, and usually they occur when the item construction deviates too far from the standard MCQ item writing guidelines, as these are the easiest within which to identify flaws and show the most predictable ones. The fact that so many of the items within the 'potential item construction issues' category are showing substantial amounts of underfit suggests that the construction issues are either confusing the students, or something about the issues may make the item exploitable; however, as this category is the broadest each item needs to be evaluated on a case by case basis to determine exactly what may be causing the issue. These sorts of issues should be easily noticeable through an item review process, and while these sorts of items do not have to be immediately removed from the assessment (as not all of them are problematic) their existence should be noted, and there should be an intention to review the performance of these items after the assessment has taken place.

Comparing the categorisation results of Chemistry IA MCQ assessments tasks to the MCQ assessment tasks of the other first-year chemistry courses analysed at The University of Adelaide there was not as large difference as was expected (the categorisation of all the first-year chemistry courses at The University of Adelaide is within Appendix 7.16). One of the issues with making this comparison is that as all of the unique items from each course were categorised it does not account for how often each of the items was asked of the students, and thus the figures are not accurate representations of the assessment tasks. The content being assessed by each course is different, but it was observed outside of that there were large similarities between Chemistry IA and Chemistry IB assessment items, and large similarities between Foundations of Chemistry IA (FoC IA) and Foundations of Chemistry IB (FoC IB) assessment items. The differences observed between those

two sets of courses was the result of the differences in the content being assessed which impacted on the construction of the items within the assessment tasks. For example, the largest difference between Chemistry IA and Chemistry IB is the large number of visual items that present the students with a chemical structure, which can be attributed to the synthetic and bio-organic chemistry topic that is taught within Chemistry IB. The largest difference between the two sets of courses was observed within the item complexity category, as both Chemistry IA and Chemistry IB had a higher proportion of multi-step and multi-mark potential items than FoC IA and FoC IB. This is likely reflective of the difference in the level that is expected from the students at the conclusion of the course, as the purpose of the FoC courses is to introduce the students to basic ideas and concepts whereas the regular chemistry courses provides more depth on the topics being taught. This is reflected within the types of items assessed, as there is a large proportion of recall/recognition items within the FoC courses; however, this is matched by Chemistry IA while almost none are present within Chemistry IB. There were also large differences observed within the potential item flaws observed in each course; however, as this is more reflective of the construction of the items and not of the trends within the course's assessment tasks, the comparison cannot be used to inform any differences between the courses. The similarities that were seen across all four chemistry courses may be due to underlying similarities between the courses, as they all assess similar chemistry topics, or it may be due to similarities in how the assessment tasks were constructed. The comparison of item categorisation between the courses can be used to inform differences between the assessment strategy employed within each course, but as the categorisation process is constructed to help inform assessors of what their assessment tasks require from the students, the comparison does not provide much meaningful information.

The categories and factors introduced and explained within this section can be used by other assessors who wish to evaluate their own assessment tasks and items. These categories and factors can be changed and altered to suit the needs of the assessors who are evaluating their item pool or an individual assessment. Thus ideally this process can be employed by any assessor to either evaluate the items before an assessment has taken place to ensure that the items within the assessment match the expectations of the assessment, or it can be used after an assessment has taken place to determine if there are any key factors that cause the items to behave in a problematic manner. Ensuring that the items used within the assessment match the expectations of the assessors is a key step in ensuring that the assessment is working as it is intended to. If all of the items behaved as they are expected to, and none of them caused any issues that were highlighted through item analysis, but the item can be answered using methods that were not intended (e.g. gaming the question), then the assessment itself is not performing its function. Thus, assessors should always be mindful that the items do not just need to be functional, they should also be constructed to align with the expectations of the assessment. Therefore, through the introduction of this process it is hoped that assessors will be able to have another way of analysing their own assessment tasks that does not rely upon a mathematical methodology and is flexible enough to suit the needs of any assessor.

3.5 Conclusion

All the data used within this research can be approximated by Normal distributions; even though some of the distributions showed statistically significant deviations from the Normal model this was due to the size of the cohorts that undertook each assessment. This allows for the application of different statistical tests to the data being used to compare between different cohorts of students.

The first step of ensuring assessment validity was to analyse the assessment task and ensure that the task is able to show the complete spread of student abilities present within the cohort and that there is a correlation between the overall results of the assessment and the results of the individual items. It is important to consider factors surrounding the assessments when evaluating these values - such as the objective of the assessment, the number of students undertaking the assessment, and the length of the assessment - as these considerations will shift the expectations of the assessment analysis. Within this research the data being utilised did show the assessments gave lower than expected student separation and reliability values which means that it likely that the student cohorts' results are not reproducible; however, this was most likely due to the small number of items used within each individual assessment task and not necessarily a result of the assessment tasks being flawed. After analysing the assessment tasks, the individual items needed to be analysed, with 12 problematic items being identified using CTT (4 minor issues, 4 potentially major issues, and 4 major issues), and 83 problematic items identified using Rasch analysis (33 minor issues, 9 potentially major issues, and 41 major issues) out of 261 unique items used in all the assessments analysed. Each of these items was further analysed using a combination of stem evaluation and distractor analysis in an attempt to determine the underlying issues. All these considerations allowed for the completion of the first objective of the research:

To assess the items used in MCQ assessments both currently and previously at The University of Adelaide in first year Chemistry courses to determine whether the performance of the students on these items is providing assessors with information that reflects the ability of the students on the content being assessed

Evaluation of both the assessment items and the tasks provided information about the performance of both the students and items that could be used to address the first research question:

Are the MCQ items used at The University of Adelaide in first-year Chemistry courses performing as they are expected to?

It was seen that the vast majority of the items within the assessment tasks were indeed performing as they were expected to, and thus they provided information about the ability of the students on the content being assessed. However, the items that were found to be problematic through this analysis are not performing as they are expected to and thus changes need to be made to those items to ensure that in the future they do not influence the validity of the assessment.

It is possible that individual items have issues that are not influential enough to cause large shifts in the results of the student cohort. These issues may instead influence the outcome of a task if an item shows a statistically significant difference in how different student cohorts perform on that item. Within this research the performance of the male and female student cohorts was compared; however, it is possible to compare between any distinct set of cohorts. The first step of this process (depending on the analytical process being used) was to ensure that the cohorts themselves do not show statistically significant differences in their abilities which would influence their results on the items. The assessment item comparison identified 27 items using CTT and 14 items using Rasch analysis to show differences in performance based upon student gender, where only 8 items were identified by both. This addressed the second objective of this research:

To compare the performance of male and female students in first year Chemistry MCQ assessments at The University of Adelaide to ensure that any difference in performance is a result of a difference

in ability and not due to factors within individual items that influence student performance based on student gender

The comparison of both the male and female student cohort abilities and their relative performance on individual items provided information to address the second research question:

Is there a significant difference in the performance of male and female students within MCQ assessments? If so, how can this be addressed?

Based on the results of this research there is no statistically significant difference between male and female student performance due to the MCQ format, and any statistically significant differences between the two cohorts could potentially be attributed to a number of considerations that need further exploration to determine the root cause. There were individual items that did show a statistically significant difference in the performance of male and female students, and these items need to be addressed, but as the majority of the items did not show any statistically significant gender differences the problem cannot be attributed to the MCQ format. It is unlikely that the root cause of the difference between male and female students' performance (where observed) will be easily identified, and thus the item needs to be improved in any way possible or removed from the assessment task to prevent those items from causing issues within future assessment tasks. Neither of these options will immediately solve the problem as improving the item may not remove the gender differences and a new item might create a new issue; however, through the continued evaluation of assessments more can be learnt and the items can be improved until there are no issues present within the assessment tasks.

All these conclusions were made using results generated using both Classical Test Theory and Rasch analysis, and while the two methodologies did give similar results there are large differences in their assumptions and the information that they generate. Most of the similarities occur when the assessment task and the student results within the task are being analysed, as in most cases the results generated by both methodologies lead to the same conclusions. The differences occur when the methodologies analyse the items individually, as the expectations of the two methods lead to different outcomes in determining which items should be considered to be problematic. This difference is also seen when the items were analysed for gender differences, and based on the results of this research it suggests that while CTT is an acceptable methodology for identifying the largest issues within an assessment task, it does not provide the information required to be able to identify more nuanced issues that can be seen through Rasch analysis.

A methodology for categorising whole assessment tasks was used as a way of ensuring that the items within an assessment task match the expectations that the assessors have of the assessment. The categories of content, taxonomy, type, presentation, process, complexity, and potential item issues were used to categorise each item, and thus the construction of the entire assessment task could be seen, as well as where any problems within the assessment were occurring. This categorisation process can either be applied before or after an assessment task has taken place to perform different functions. If the items are categorised before the assessment task is used it will provide an overview of the construction of the assessment task, and thus can be used to ensure that the assessors have created an assessment that matches its purpose. Depending upon the purpose of the assessment this may be seen within different facets of the categorisation process; for example, if the task is intended to assess the students only in one particular subject and only their basic recall of the relative information, then content and type are the most important categories to consider.

Categorising the assessment task after the assessment has taken place can be used in conjunction with item analysis to determine if there are any specific areas that are causing issues for the students and what potential factors may causing this, based on what other categories those items lie within. It is also possible to use an item's categorisation to replace it with a similar style of item if an item needs to be removed from the assessment task for any reason.

Chapter 4: Assessments as Comparable Measures of Performance

4.1 Section Outline

4.1.1 Research Questions

Once the data and the items have been validated and any significant areas of concern are identified, it is possible to use the assessment results to make several comparisons using students and items that are common across multiple assessments. In this research, the assessments were used to compare how students perform in the test-retest assessment structure that is used in first year Chemistry courses at The University of Adelaide. This is possible due to the students being given two opportunities to undertake a similar assessment task with only their best result considered for their course outcome. Students' first opportunity is during the semester where there are two lecture tests; the first takes place halfway through the semester and the second at the end of the semester. The second opportunity is during the final exam where an hour is allocated specifically for the redeemable MCQ assessment task. The results of the students during the semester were compared to their results within the redeemable section of the final exam to determine if there are significant changes in their performance, addressing one of the research questions:

Do students show differences in their performance in MCQ assessments at different points in a semester? If so, how?

Similarly, since the first year Chemistry MCQ assessments show a large amount of commonality in the items they use between years it is possible to use those items to link the assessments in such a way that the results of student cohorts over multiple years can be compared. This comparison enables the changes in the student cohort over multiple years to be determined, which can be used to determine if the average ability of the student cohort is significantly changing over time. This allows for another of the research questions to be addressed:

Do student cohorts show differences in performance over multiple years? If so, how?

Finally, in the same way that items can be used to link assessments, there is the potential to use students that are common to multiple assessments to link the assessments. Thus, there is the potential to compare between courses from different disciplinary areas if there are students that are enrolled into both. A large amount of student overlap is seen between first year Chemistry and first year Biology courses at The University of Adelaide, giving the potential to compare these two courses. Attempting to make a comparison between these courses will address another one of the research questions:

Is it possible to compare student results across multiple courses from different disciplinary areas? If so, do students show similar performance across multiple courses?

All these research questions will be addressed using items and students that are common to multiple assessments, making this comparison uniquely suited to MCQ assessments due to their objective nature and due to the assessments having had very little adjustments over multiple years.

4.1.2 Project Objectives

Being able to compare student ability and item difficulty across different assessments and courses addresses one of the objectives of this research:

To compare item and student performance within first year Chemistry assessments over the period of a semester, across multiple years, and against Biology courses using MCQ assessments undertaken at The University of Adelaide to determine if there are any differences in performance, and if these changes are a result of the items or the students

Upon completion of this objective it will be known if there are significant changes in student ability within the semester, which will inform whether students perform better after learning all course material or if they remember the information 'better' when the assessment is undertaken shortly after that content is introduced. The objective will also give information on whether yearly student cohorts show significant differences in their average ability, which can inform the expectations of student outcomes in future years. Having the ability to compare between courses provides information on the relative difficulty of those courses and if students have similar ability across courses; thus, it can be used to address assumptions made about those two considerations.

4.2 Effects of Test-Retest Assessment on Student Performance

4.2.1 Assumptions and Methodology

The two MCQ assessments tasks undertaken within lecture tests for all first year Chemistry courses at The University of Adelaide are redeemable within the final exam, as this provides the students the opportunity to improve upon their performance or sit an assessment task they may have missed during the semester. Thus, it is possible that students undertake the same, or a very similar, MCQ assessment task twice in the same semester. It is also possible that students are only undertaking the assessment task on one of the two occasions that the assessment is offered, as the students may decide that they only want to sit the assessment task on one occasion for a variety of potential reasons. This gives the opportunity to not only compare how the performance of the students changes throughout the semester and relative to the rest of the student cohort, but also how students who only undertook the assessment on one occasion perform. The results of the students can only be compared when there is an adequate number of items that are utilised in both the lecture tests and final exam. If this is not the case, then there is no way to be sure that the two assessments are of equal difficulty, and thus students scoring lower on one of the assessment tasks may be due to the items being asked rather than differences in student performance, which is why only 8 comparisons could be made within this research. This is a larger issue in CTT than it is in Rasch analysis, as CTT has no way of accounting for items that are not present within both assessments. This means that any items that are not present in both assessments need to be removed from the analysis when using CTT, which leads to questions about the validity of the analysis as it is no longer representative of the results from the entire assessment. Rasch analysis only requires enough shared items to anchor the two assessments together (see Section 2.6.3), which usually can be done using a minimum of five items that are shared across both assessments.[309,330] If the two assessments can be anchored using those items it means that both the student and item measures produced from both assessments will be expressed within the same logit scale, making them comparable measures. Once the item difficulty and student ability measures are shown to be comparable across assessments it is simply a matter of determining whether there is a statistical difference between the item difficulty and student ability measures produced in each assessment.

It cannot be assumed that a student's approach to both the lecture tests and the final exam MCQ assessment tasks will be the same, and thus comparing performances over the two assessments gives information about the behaviour of the students as well as identifying differences in their performance. It is expected that the students will have at least taken slightly different approaches to those tasks, as the student's study pattern before an exam and before a MCQ lecture test is likely influenced by the breadth of knowledge being assessed, the amount of time they have to prepare, and the different assessment formats that are present. It is not an issue if the student's study approach is different between the assessment tasks, as regardless the assessment task will represent the student's competency within the material being assessed; this only becomes an issue when the differences in student approaches have the potential to influence their results. An example of this would be if a student did not study at all for the lecture test with the intention of either treating the assessment purely as practice, or potentially identifying how much information they have retained without any revision. If that same student then approached the exam by meticulously revising all of the content that had been taught throughout the course then it should be expected that the student is going to perform better on the exam than they did in the lecture tests. However, part of the reason that the students are given the chance to redeem their marks within the final assessment is because they are expected to increase their knowledge throughout the semester, and thus this gives the student an opportunity to exhibit how much they have learnt. In contrast to this example it is also possible that the students may not improve their results within the final exam. This could be the result of a variety of factors, but one possibility is that the students revised before the lecture test with the intention of overperforming. These students may then attempt the redeemable section within the exam not with the intention of performing better than they did on the lecture test, but rather because there is no negative impact for them to at least attempt the assessment with the possibility that they may improve upon their previous result.

4.2.2 Classical Test Theory Analysis

The first step in comparing the two assessment tasks is to identify the items that were utilised on both assessment tasks, as only assessment tasks with a substantial number of shared items can be compared. This is because CTT has no way to compare the results of students on different items within the assessment tasks, and as they have the potential to influence the results of the students they cannot be included within the comparison. Thus, it can be seen by viewing Table 27 that only Chemistry IA and Chemistry IB had assessment tasks that were appropriate for comparing the lecture tests and the final exam, and thus the assessments within both Foundations courses were unable to be compared.

Table 27: The Number of Items that are shared between the Lecture Tests and the Final Exam within Each of the Courses Analysed within this Research

	2012	**2013**	**2014**	**2015**
Chemistry IA	19	19	19	18
Chemistry IB	24	24	24	25
Foundations of Chemistry IA	0	0	0	0
Foundations of Chemistry IB	0	1	3	3

The reason that items used in entire assessment tasks were not the same in any of these cases is because some items either underwent minor changes or were completely removed and replaced within the final exam. The results from the items that were not shared had to be discarded for the purposes of this comparison, as it cannot be assumed that the students' results in any of those items

is mirrored within the item that replaced it. The item difficulty of these items then needs to be compared across the two different assessment tasks to ensure that the item performs similarly in both tasks.

To compare between the item difficulties (P) of the shared items used within the lecture tests and the final exam, a paired sample t-test was used as this compares how the mean of the item difficulties has changed between the two instances. It is expected that there is no difference between the mean item difficulties of the items utilised in both assessment tasks, as otherwise it implies that the items are behaving differently in the different assessment tasks. The results of this comparisons can be seen in Table 28.

Table 28: Comparison of Item Difficulties of Items used within both the Lecture Test and the Redeemable Exam to Determine if there is a Significant Shift in Mean Item Difficulty. Highlighted Cells indicate Observation of a Statistically Significant Difference

	Year	Assessment Task	Mean	S.D.	d.f.	*p*-value
Chemistry IA	**2012**	**Lecture Test**	0.584	0.177	18	0.084
		Final Exam	0.617	0.188		
	2013	**Lecture Test**	0.600	0.181	18	0.705
		Final Exam	0.611	0.181		
	2014	**Lecture Test**	0.561	0.185	18	0.425
		Final Exam	0.584	0.148		
	2015	**Lecture Test**	0.584	0.192	17	0.543
		Final Exam	0.599	0.154		
Chemistry IB	**2012**	**Lecture Test**	0.544	0.130	23	0.005
		Final Exam	0.576	0.140		
	2013	**Lecture Test**	0.567	0.136	23	0.001
		Final Exam	0.611	0.136		
	2014	**Lecture Test**	0.577	0.137	23	0.008
		Final Exam	0.608	0.131		
	2015	**Lecture Test**	0.572	0.156	24	0.001
		Final Exam	0.611	0.139		

There is no statistically significant difference observed between the mean item difficulty within Chemistry IA lecture test and final exam MCQ assessment tasks; however, there is a statistically significant difference observed between the mean item difficulties in all Chemistry IB lecture test and final exam MCQ assessment tasks. It should be noted that the value of the mean item difficulty increased within the redeemable section of the final exam in comparison to the mean item difficulty within the lecture tests (meaning the students found the redeemable section of the final exam to be easier than the lecture tests); however, the difference was only ever significant ($p<0.05$) in Chemistry IB. Item difficulty generated using CTT represents the percentage of the student cohort who selected the correct response on the item. This means that the item difficulty is dependent upon the ability of the student cohort that sat the assessment, and thus if the ability of the student cohort shifts then the item difficulty will also shift as a result. For that reason comparing the item difficulties as a way of ensuring that the performance of the items has not changed may not function as intended; however, as this is the only way to track potential changes in item performance using CTT, there is no alternative but to use it despite the potential for it to be influenced by changes in student performance. An increase in the value of the item difficulty is expected if it is thought that

the students' performance within the assessment task is also improving. Thus, even though these differences are statistically significant they may not indicate significant changes in item performance, but rather they may indicate changes in student performance. The deviations from the mean item difficulty in the different assessment tasks ranges from 0.01 – 0.05 across all the assessments compared (as seen within Table 28). This effectively represents a difference of 1 -5% of the student cohort obtaining the correct answer more often than they did within the lecture tests. Thus, despite being classified as statistically significant it is reasonable to theorise that the items are performing similarly in both assessments, and the changes are due to the dependence of the item difficulty on the student ability. If there is no significant difference in student ability then this assumption needs to be re-evaluated, as it means that the statistical significance observed may not be accounted for by differences within the student cohort. This would therefore suggest that the statistically significant difference observed is either the result of shifts within item difficulty that are not reflected within the student ability (which may be due to students who only undertake the assessment task on one occasion), or how the students perform on the items changes between assessment tasks independent of their ability.

Comparing the mean raw score of the students over multiple assessments is done in the same way, using a paired sample t-test to compare the raw scores that each student obtained on the same items on two different assessment tasks (lecture test and final exam). When considering the results of each individual student a raw score of 0 had to be treated as if the student had not undertaken the assessment task (there is no way to differentiate between a student obtaining a score of 0 and a student who did not undertake the assessment task), and hence they were removed from the comparison. This is another reason why removing items that were not utilised in both assessment tasks can complicate the comparison, as this minimises the number of items present within the comparison and hence increases the chances that students need to be removed from the analysis. It should be noted that only the students who sat both lecture tests were incorporated within the comparisons being made, as there were a number of students who only sat the first lecture test and no other assessment. There may be a variety of reasons why a student may only undertake one of the two lecture tests; however, whatever the reason, as there is no way within this analysis to account for a student only undertaking only one of the two lecture tests they needed to be removed. The results of the comparison of the student cohort mean raw scores in both assessment tasks can be seen in Table 29.

Table 29: Comparison of the Student Cohort Mean Raw Score on Shared Items between the Lecture Test and the Redeemable Exam Section to Observe Changes in Student Performance. Highlighted Cells indicate Observation of a Statistically Significant Difference

		Assessment	Mean Raw Score	Items	d.f.	*p*-value
Chemistry IA	**2012**	**Lecture**	11.65	19	397	<<0.001
		Exam	12.16			
	2013	**Lecture**	11.88	19	369	0.002
		Exam	12.30			
	2014	**Lecture**	10.76	19	399	<<0.001
		Exam	11.49			
	2015	**Lecture**	10.48	18	408	<<0.001
		Exam	11.32			
Chemistry IB	**2012**	**Lecture**	13.04	24	322	<<0.001
		Exam	14.64			
	2013	**Lecture**	13.63	24	315	<<0.001
		Exam	15.51			
	2014	**Lecture**	13.54	24	341	<<0.001
		Exam	15.23			
	2015	**Lecture**	14.30	25	359	<<0.001
		Exam	15.94			

These results show that there is a consistent statistically significant increase in the performance of the students on the common items that were utilised in both the lecture test and the final exam. Like previous instances of significance testing on the student cohort, there is a risk that significance is only seen because of small differences within the cohort that appear significant due to the large sample size. Thus, the mean results of the students were given within Table 29 to provide a way to compare the difference between the mean raw scores of the students on both assessment tasks. Using the data from the table, when there is a shift in the student performance within the redeemable section of the final exam it usually represents a shift of +0.5-1 mark across all of the student cohorts compared in Chemistry IA assessment tasks and +1-2 marks across all the student cohorts compared in Chemistry IB. This shows that even though there is a statistically significant change in the performance of the students the actual size of the difference itself may not be large. However, the issue with comparing the mean results of the students is that it suggests that the entire cohort of students is shifting in the same way (i.e. that all the students improve their performance in the redeemable section of the final exam or vice versa). A way to visualise how student performance changes is by using a cross-plot of the raw scores obtained by the students in each assessment task. An example of such a cross plot for Chemistry IA (2012) can be seen in Figure 27 (see Appendix 7.17 for the cross-plot of every assessment analysed). Here it can be seen that many students are not consistent in their performances across both assessment tasks and although the overall performance of the student cohort improves not all the individual students do.

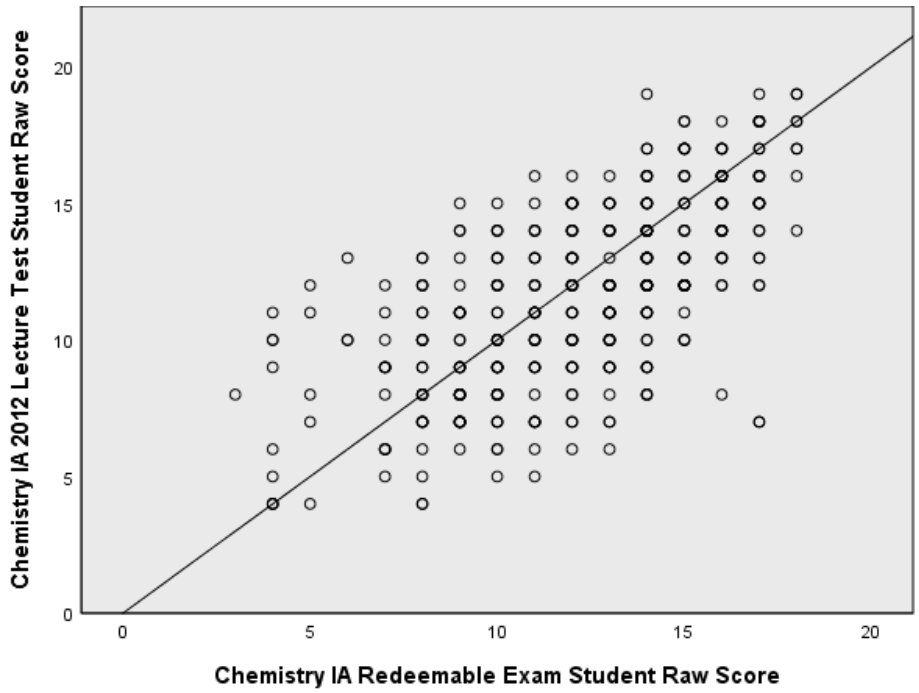

Figure 27: Cross-plot Comparison of the Student's Raw Score Obtained Using the Same Items in Two Different Assessments within Chemistry IA in 2012 to View Changes in the Performance of the Student Cohort

If a student's performance lies on the linear fit it means that their performance within the lecture test matches their performance within the redeemable section of the final exam. If a student's performance is below the linear fit it means that they have performed better in the redeemable section of the final exam in comparison to the lecture test. If their performance is above the fitted line, then they performed better within the lecture tests than the redeemable section of the final exam. There is a large amount of overlap in student scores due to the limited number of possible scores within both assessment tasks, which is far exceeded by the number of students within the cohorts who undertook these assessments. This is why some of the data points within the cross-plot appear darker than others, as this gives some indication of the number of students who obtained that combination of scores; however, this information cannot accurately be determined from the cross-plot. While a statistically significant difference in student performance was observed the fact that not all the students improved their raw scores raises questions about how the significant improvement changes the distribution of the students' results. This kind of result is repeated throughout all the comparisons that were made between the lecture tests and the redeemable section of the exam (see Appendix 7.17). This would tend to indicate that while there is a statistically significant improvement in the performance of the students from the lecture tests to the redeemable section of the final exam for all the cohorts analysed, this change is not consistently observed across the entire student cohort in any of the assessment tasks.

Another consideration when analysing the student cohort distribution is how the students who only undertake one assessment perform, as there is the potential that the improvement seen within the redeemable section of the exam may be due to factors that are unrelated to having a second attempt at the items. This could include factors such as: extensive exam preparation, completion of the course material, and having more experience applying the course content. The two best ways to view a comparison of the student distribution is using a box plot and a histogram which can show how the students compare in each sitting of the assessments, and also how the students who only

undertook one of the assessments compare to the students who undertook both. These graphs can be seen for the Chemistry IA 2012 student cohort in Figure 28 and Figure 29.

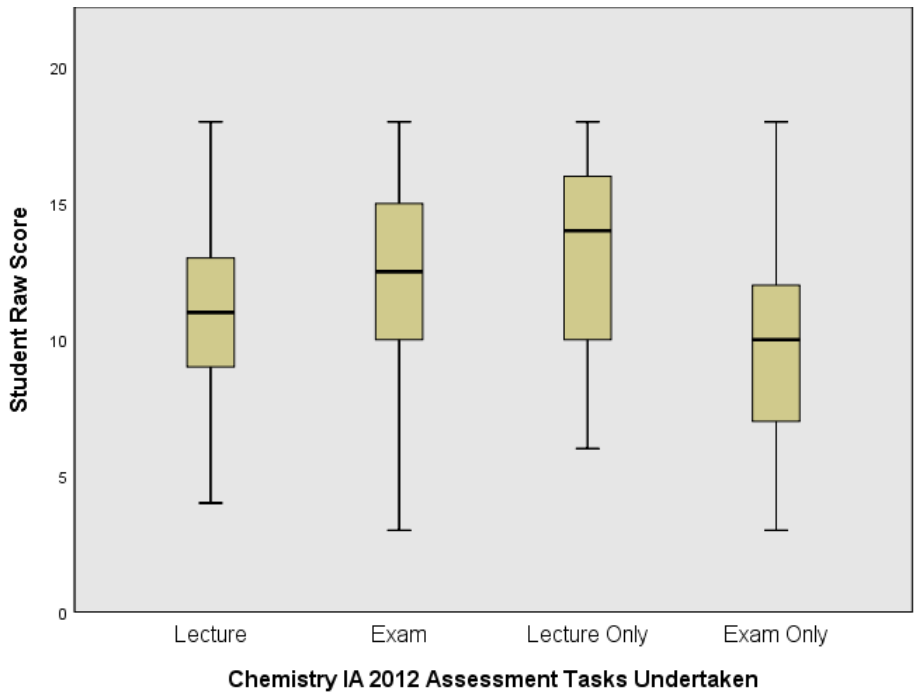

Figure 28: Boxplot Distribution of the Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA 2012, showing the Ability of the Students who undertook both Test and Retest and those who undertook one Assessment Task

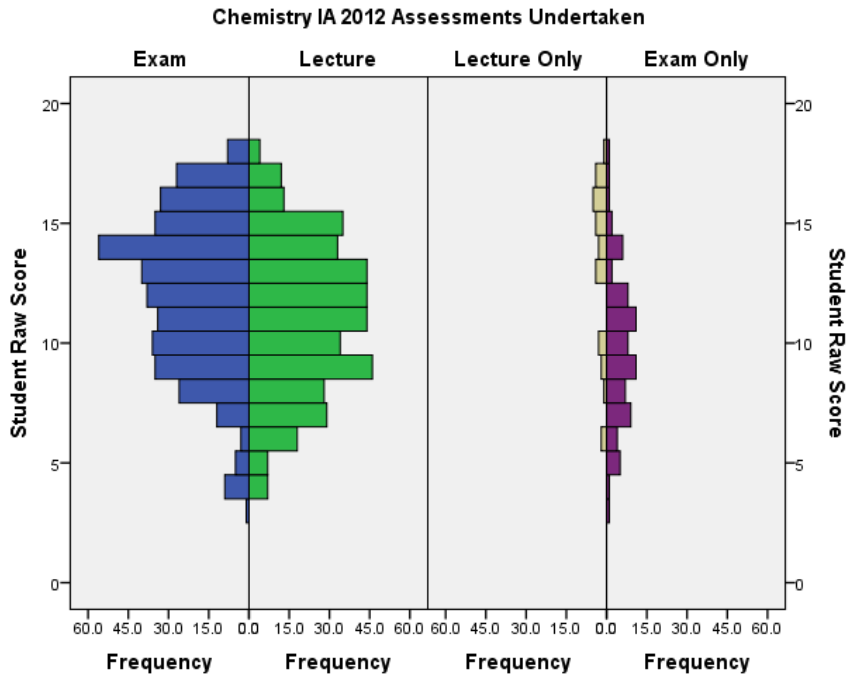

Figure 29: Histogram Distribution of the Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA 2012, showing the Ability of the Students who undertook both Test and Retest and those who undertook one Assessment Task

The figures illustrate the shift in performance between the lecture test and the redeemable section in the final exam, as it can be observed that within the redeemable section of the final exam the scores of the students on the items utilised within both assessments was higher. This shift within the student cohort results is what causes the statistically significant difference between the student cohort's mean raw score within the lecture test compared to their raw score within the redeemable section of the final exam.

Based on these figures it can be theorised that the students who choose to only sit the lecture tests do so because they either did not believe that they could perform better within the redeemable exam due to the high marks that they obtained, or they were satisfied with their original result, although the vast majority of students choose to sit both assessments (398 of the 506 students who undertook Chemistry IA in 2012). The trends seen here were consistent across the other seven analyses that were undertaken for the other years and course being analysed (see Appendix 7.18 for the other analysis results). This would appear to indicate that students who only undertook the lecture tests tend to achieve above average results, while the students who only sat the redeemable section of the final exam tend to achieve average to below average results in comparison to the student cohort who undertook both assessment tasks.

The tendency of students who only undertook the redeemable section of the exam to obtain lower results is likely because this is an indicator of the students' level of engagement and attitude toward the course. There are multiple valid reasons as to why students may not undertake the lecture tests; however, the students who choose not to undertake the assessments when they could have may do so because they are disengaged with the course at the point in time that the assessment task takes place. There are a variety of reasons that the students disengage with a course, but it does not mean that the student is disengaged with their learning in general. For example, Chemistry IA is a required course for many different programs that require the students to complete first year Chemistry. It is reasonable that some of these students may lack some motivation for Chemistry IA as they are only undertaking the course as it is a requirement of their program, and thus potentially their objective is to simply pass the course and focus on other aspects of their program. A further comparison of the students can be made by identifying the students that show improvement, deterioration, or have no change in their assessment results, as well as the students who only undertook the assessment task on one occasion. This comparison can be seen in Table 30.

Table 30: The Raw Score Average Result from Shared Items within Chemistry IA Assessments from 2012 and how they Shift with Different Student Performance

19 Items Shared		Student Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	**Cohort**	506	11.07	3.17	12.16	3.28	
	Lecture Only	29	13.38	3.38			
	Exam Only	79			9.65	3.26	
Test-Retest Changes	**Increase**	247	10.32	2.98	13.04	2.78	2.73
	Decrease	105	12.33	2.76	10.03	3.16	-2.30
	No Change	46	12.26	0.08	12.26	0.08	

The average increase in the performance of the students that improved within the redeemable section of the final exam compared to the lecture tests is closely matched by the average decrease in performance of the students whose marks decreased within the redeemable section of the final

exam. This suggests that the significant increase in student performance observed previously is a result of the number of students improving, and not due to the size of their improvement. However, the fact that more students are improving their result within the redeemable section of the final exam compared to their performance within the lecture tests is still of interest. Thus, giving the student a second opportunity to undertake the assessment does allow some students to clearly demonstrate improvements within their ability, while not negatively impacting students who perform better shortly after they were introduced to the concepts being assessed.

The large difference observed in student performance when they only sat one assessment supports the theories previously discussed when comparing the student distributions. These trends are also mirrored within the comparisons made between the two assessments within the other years and courses being analysed (as can be seen within Appendix 7.19). The results of other analyses also showed that the average change in performance is matched closely with the students that both improve and degrade, and thus is the number of students that improve that causes the statistically significant difference within the results of the lecture test and the redeemable section of the final exam.

The large difference in the performance of the students who only attempted one of the two assessment tasks can also be seen within the table. It was observed that the highest average marks comes from the students who only undertake the lecture tests, while the lowest average marks comes from the students who only sit the redeemable exam. This was also a result that was seen within the other comparisons that were made within Chemistry IA and Chemistry IB (see Appendix 7.19), and this amount of consistency suggests that there may be some commonality in the reasons that the student cohorts behave the way that they do.

The fact that the students who only sat the lecture tests had the highest average result, and the students who only sat the redeemable exam had the lowest average result does suggest that undertaking both assessment tasks gives the students the best chance of success; however, these groups contained far less students than the number that undertook both assessment tasks. This may indicate that the results of these student cohorts may be skewed in one direction due to the smaller sample size, which means that individual students have more impact on the observations. Using information on how the students who only undertook one assessment task performed relative to students who undertook both can be used to determine if the results are influenced by outliers. The comparisons of the students who only undertook the lecture tests compared to the students who undertook both assessment tasks can be seen within Figure 30. Similarly, the comparison of the students who only undertook the redeemable section of the final exam to the students who undertook both assessment tasks can be seen within Figure 31 (see Appendix7.20 for the distributions in the other assessments analysed).

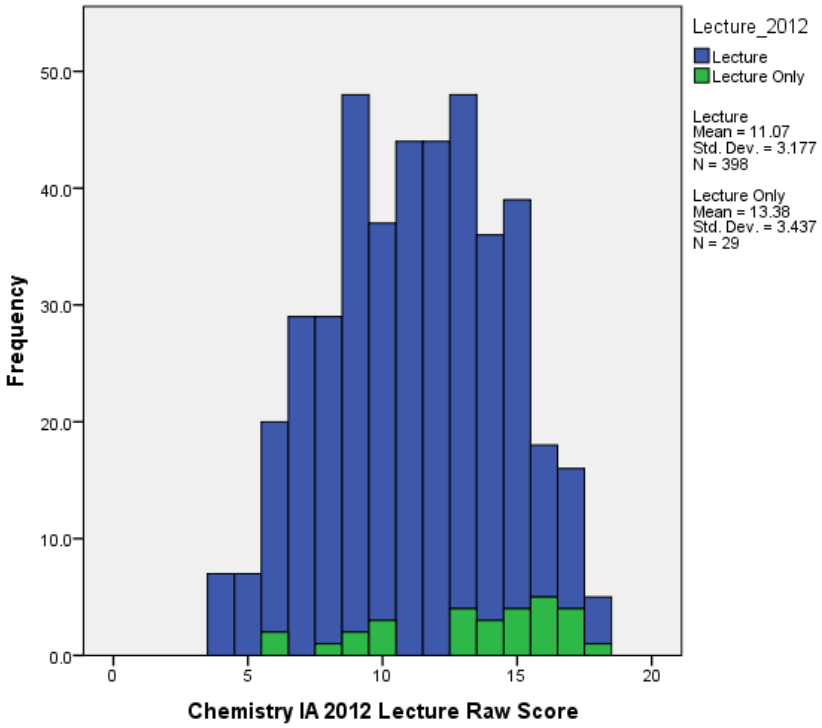

Figure 30: The Results of the Students on Shared Items within both Lecture Tests in Chemistry IA in 2012 Comparing Students who only undertook the Lecture Test (Lecture Only) to those who undertook both Assessment Tasks (Lecture)

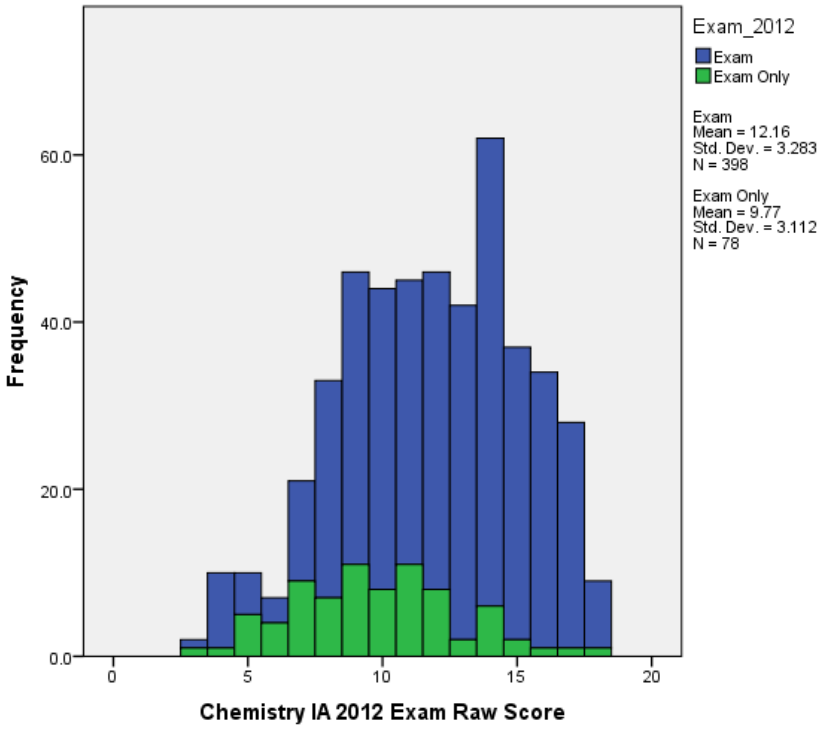

Figure 31: The Results of the Students on Shared Items within the Redeemable Section of the Exam in Chemistry IA in 2012 Comparing Students who only undertook the Redeemable Section within the Final Exam (Exam Only) to those who undertook both Assessment Tasks (Exam)

These figures clearly show that while there is a spread of results for the students who only undertook the assessment once, the students who only undertook the lecture tests are often amongst the highest performers (21 of 29 students scoring above the average of the student cohort who undertook both assessment tasks), and the students who only undertook the redeemable exam are often amongst the lowest performers (58 of the 79 students scored below the average of the student cohort who undertook both assessment tasks). These results are likely related to the motivation of the students; however, as discussed previously there are many reasons why the students may only undertake the assessment once and without discussing it with the students it is impossible to know their reasoning.

The results of the comparisons from all the assessments show a similar trend, which suggests that allowing the students to sit the assessment on multiple occasions gives them the best opportunity to display their knowledge and ability. Redeemable assessments are not always appropriate within courses as they do add several complications including: finding the time for an additional assessment, influence it has on student behaviour, and maintaining the validity of the assessment. Therefore, while the use of redeemable assessments has been shown to be beneficial for the results of the students this does not mean that it is appropriate within all courses, and for all assessments. For example, it is not feasible for the entirety of the final exam to be redeemable as it is used to determine the students' competency across everything that was taught throughout the course. Nor is it feasible for all assessment formats to be redeemable, as having the students undertake assessment tasks such as practical or oral examinations on multiple occasions would be highly time-consuming for both the student and the assessor. Thus, the most feasible redeemable assessment format is the MCQ format, as not only is it the least time-consuming for the students to complete, but also for the assessors to mark.

The greatest concern for an assessor when offering redeemable MCQ assessments is maintaining their validity, as if the same items are reused there is the potential that students may simply memorise those items in preparation for the redeemable assessment. However, it is highly unlikely that the students will be able to memorise the entire assessment (especially if the questions are only available to students for the duration of the assessment task and not published elsewhere, as is the case for assessment tasks utilised in this research) and if this is a large concern the items can either be changed in subtle ways that influence the answer, such as changing the numbers for calculations. It may be possible to replace some of the items between the two assessment tasks; however, this should be done with caution, as unless the items are replaced with new items that assess similar or the same content it may negatively impact the validity of the redeemable assessments. The new items can influence the validity of the redeemable assessment tasks in two ways: they may either change the difficulty level of the assessment task, thus causing the two assessment tasks to not be equivalent to each other, or they may introduce problematic items that cause issues unrelated to the content being assessed. Therefore, even though it might be thought that the student improvement observed within this research may be the result of students attempting to memorise the assessment task, if a redeemable format is utilised it should include as many shared items as possible to minimise other validity concerns.

As CTT was used to determine this result it meant that large assumptions needed to be made about the comparability of the results of the two assessments tasks. CTT has no way to ensure that the assessment task and items are behaving in the same way in both undertakings due to their dependency on the student cohort. Thus, any changes observed within the task and items must be treated as though they were caused by the student cohort even though there is no evidence for that

assumption. The other large concession that needs to be made for CTT is that only the results of items that are utilised within both assessment tasks can be compared, and thus it has to be assumed that the results on those items are reflective of the students' performance across the entire assessment task. Even though these large assumptions needed to be made, there was a high level of consistency within the results observed, which suggests that the trends observed are not the result of random variations caused by the methodology but rather a trend that should be expected when utilising redeemable assessment tasks.

4.2.3 Rasch Analysis

The comparison of test-retest student performance can also be made using the Rasch model, where instead of using the raw scores the Rasch ability measures can be used to compare across the lecture tests and the redeemable section of the final exam. The greatest advantage to using the Rasch model is that unlike in CTT the item difficulty measures are independent of the student cohort, and the student ability measures are independent of the items used within the assessment tasks. This has two significant effects on the analysis: the first is that the entire assessment task can be used (not just the items utilised within both assessments) as long as the measures are placed on the same logit scale and the same ability is being measured by both assessment tasks; and the second is it means that when comparing how the item difficulties change between assessment tasks there is no need to consider if the student cohort is influencing the results. This means that Rasch analysis can be used to obtain a comparison of student performance across the entirety of the assessment tasks rather than just the items utilised in both tasks. There is still a requirement for the assessments to contain several shared items to ensure that the measures can be placed on the same scale, which meant that only two of the chemistry courses could be compared, as only those two courses had enough items to link the two assessment tasks (see Table 27 for the number of shared items within each course analysed). The lecture tests and the redeemable section of the final exam were linked by stacking the items present within both assessment tasks: students were represented by rows and appeared twice (once for each assessment task undertaken [Lecture Test 1 and Lecture Test 2 were placed together]) and the items were represented by columns where each unique item only had one column regardless of which assessment task it was used in (see Section 2.6.3 for details on linking assessments). This ensured that all the item difficulties and student abilities were generated on the same scale. The process of this analysis follows the same logic that was used previously within the CTT analysis, where first the item difficulty was compared to ensure that the items were performing the same way in both assessment tasks. While in CTT the item difficulties were compared to ensure that they did not influence the student cohort, in Rasch analysis they are compared to ensure that the items can be used to link the assessments without introducing any bias into the logit scale. This is done using a paired-sample t-test to compare the mean item difficulty of the items that are used within both assessment tasks, where the null hypothesis is that the item difficulties will be the same within both assessment tasks after they have been linked. The results of this analysis can be seen in Table 31, which shows that there is no significant difference between the item difficulty measures from the lecture tests and the redeemable section of the final exam.

Table 31: Comparison of the Mean Item Difficulty Measures of Shared Items from the Lecture Tests and the Redeemable Section of the final Exam to Determine if there is a Significant Difference in Item Performance. Highlighted Cells indicate Observation of a Statistically Significant Difference

	Year	Assessment Task	Mean Item Difficulty	S.D.	d.f.	*p*-value
Chemistry IA	**2012**	**Lecture Test**	-0.090	0.951	18	0.984
		Final Exam	-0.092	1.092		
	2013	**Lecture Test**	-0.116	1.033	18	0.970
		Final Exam	-0.122	1.034		
	2014	**Lecture Test**	0.120	0.953	18	0.968
		Final Exam	0.114	0.777		
	2015	**Lecture Test**	0.042	0.994	17	0.997
		Final Exam	0.043	0.784		
Chemistry IB	**2012**	**Lecture Test**	0.037	0.663	23	0.857
		Final Exam	0.027	0.730		
	2013	**Lecture Test**	0.048	0.684	23	0.966
		Final Exam	0.045	0.732		
	2014	**Lecture Test**	-0.006	0.722	23	0.943
		Final Exam	-0.010	0.708		
	2015	**Lecture Test**	0.019	0.850	24	0.983
		Final Exam	0.018	0.814		

Even though there were 30 items within each assessment task, only the items that were present within both assessments were compared, as there is no reason to believe that the items that were not utilised in both assessment tasks would have comparable item difficulties. Despite this, the student abilities did not need to be adjusted due to the independence of the student ability measures from the items. It is also possible to view how each individual item changes if there is a concern that the mean result is masking shifts within the performance of individual items. This can be done using a differential item functioning (DIF) plot, which shows the differences in the item difficulty when asked within the lecture test and the redeemable exam, an example of which can be seen in Figure 32.

Based on Figure 32 there are several items whose DIF size may be large enough to be statistically significantly different between their difficulty within the lecture tests and the redeemable section of the final exam (a difference in item difficulty greater than 0.50 is statistically significant). In this research, these differences were not seen as a concern since student ability and item difficulty are independent of each other. This means that so long as the student and item measures generated from the assessment tasks are within the same logit scale it does not matter if the item difficulty of some of the items are statistically significantly different from each other as this will not impact the comparison of the student ability measures. Thus, for the rest of the assessments, the paired-sample t-test was a sufficient method for comparing the item difficulty measures of the shared items to ensure that the linking of the assessments does not contain any unrelated bias.

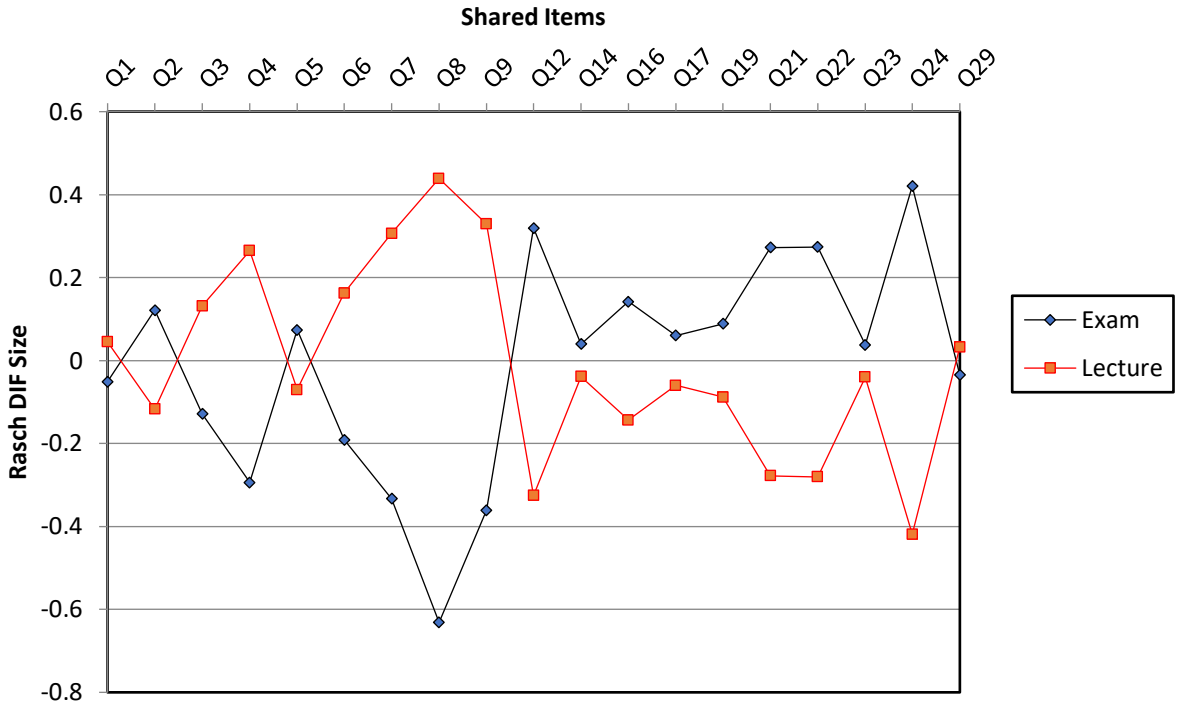

Figure 32: DIF Plot Showing the Difference in Shared Item Difficulty Measures from the Lecture Test and Redeemable Exam Within Chemistry IA in 2012

The use of the Rasch model also meant that there was no requirement for the students to have undertaken every item within any of the assessment tasks, as a comparable ability measure can be generated despite the differences in the number of items answered. Despite this, students who obtained the minimum score of zero were still removed from the analysis as it remains unclear if they truly obtained a minimum score or if they did not attempt the assessment task. The results of the comparison between student ability measures generated using their results on both lecture tests and their ability based on their results within the redeemable section of the final exam can be seen in Table 32, which is a paired sample t-test comparison with the null hypothesis that the average student ability within both assessment tasks will be equivalent to each other.

This shows that in every comparison the students performed better in the redeemable exam than they did within the lecture tests. This implies that the mean student ability increases throughout the semester, and as was discussed within the CTT analysis section, even though there is a statistically significant increase in the mean student ability this does not mean that all of the students within the cohort improve within the redeemable section of the final exam. To view how the performance of individual students are changing, a scatterplot can be used. This comparative plot can be seen in Figure 33.

Table 32: Paired Sample t-test Comparison of the Rasch Student Ability Measures on the Lecture Test and the Redeemable Exam to Observe Changes in Student Performance. Highlighted Cells indicate Observation of a Statistically Significant Difference

		Student Category	Mean Student Ability	S.D.	d.f.	*p*-value
Chemistry IA	**2012**	**Cohort Lecture**	0.255	0.899	452	<<0.001
		Cohort Exam	0.580	1.073		
	2013	**Cohort Lecture**	0.412	0.988	438	<<0.001
		Cohort Exam	0.561	1.086		
	2014	**Cohort Lecture**	0.393	0.929	460	<<0.001
		Cohort Exam	0.600	1.079		
	2015	**Cohort Lecture**	0.394	0.931	484	<<0.001
		Cohort Exam	0.596	0.969		
Chemistry IB	**2012**	**Cohort Lecture**	0.220	0.926	379	<<0.001
		Cohort Exam	0.525	0.960		
	2013	**Cohort Lecture**	0.317	0.871	373	<<0.001
		Cohort Exam	0.751	1.084		
	2014	**Cohort Lecture**	0.304	0.937	407	<<0.001
		Cohort Exam	0.652	1.068		
	2015	**Cohort Lecture**	0.341	0.945	422	<<0.001
		Cohort Exam	0.772	1.163		

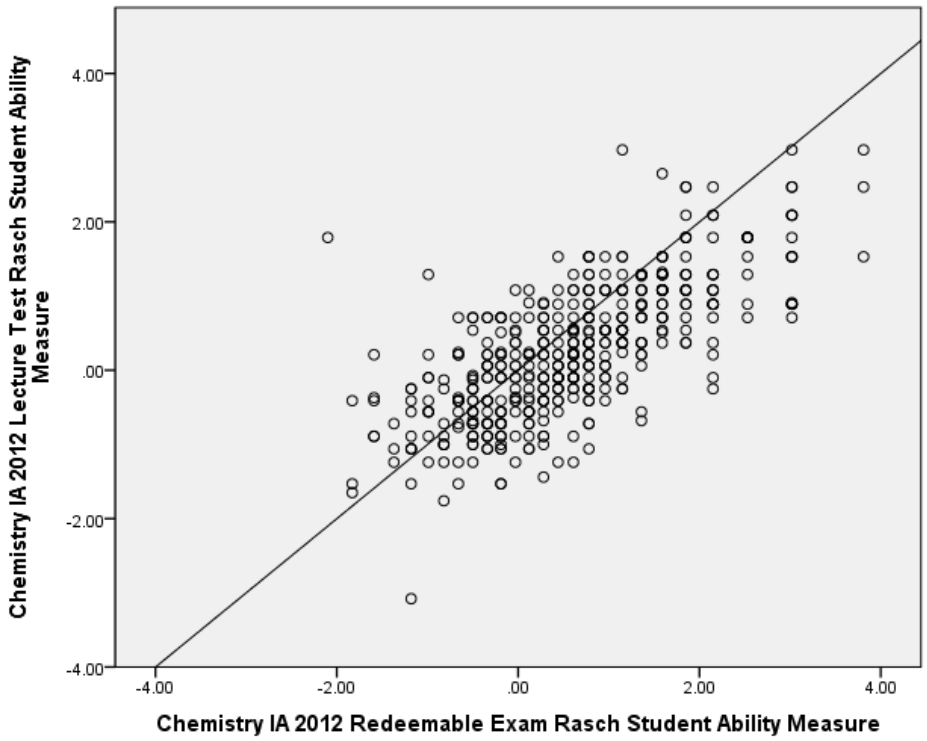

Figure 33: Scatterplot Comparison of the Student's Ability Measure Obtained Within the Shared MCQ Assessment Items from Chemistry IA in 2012 to Changes in the Performance of the Student Cohort

It can be seen within the figure that while most of the students are improving between assessments (lying below the line means a higher ability was measured within the redeemable exam assessment),

not all of them are. This indicates that there is inconsistent behaviour across the student cohort, and thus either the students are taking different approaches, or some students may not be retaining the information after it was taught. The same trend is observed when viewing the scatterplot of other comparisons that were made within this research (see Appendix 7.21 for the scatterplots of all assessments compared), showing that the majority of the students have a higher ability measure within the redeemable section of the final exam but not all of them follow this trend.

Despite this inconsistency in the performance of the students, there was a statistically significant difference observed with the student ability measure between the lecture tests and the redeemable section of the final exam in every comparison made within this research. To see how the distribution of the student ability measures are changing between the lecture tests and the redeemable section of the final exam it is possible to use a boxplot and/or a histogram. Examples of these plots showing the difference within the mean ability measure of the students can be seen in Figure 34 and Figure 35.

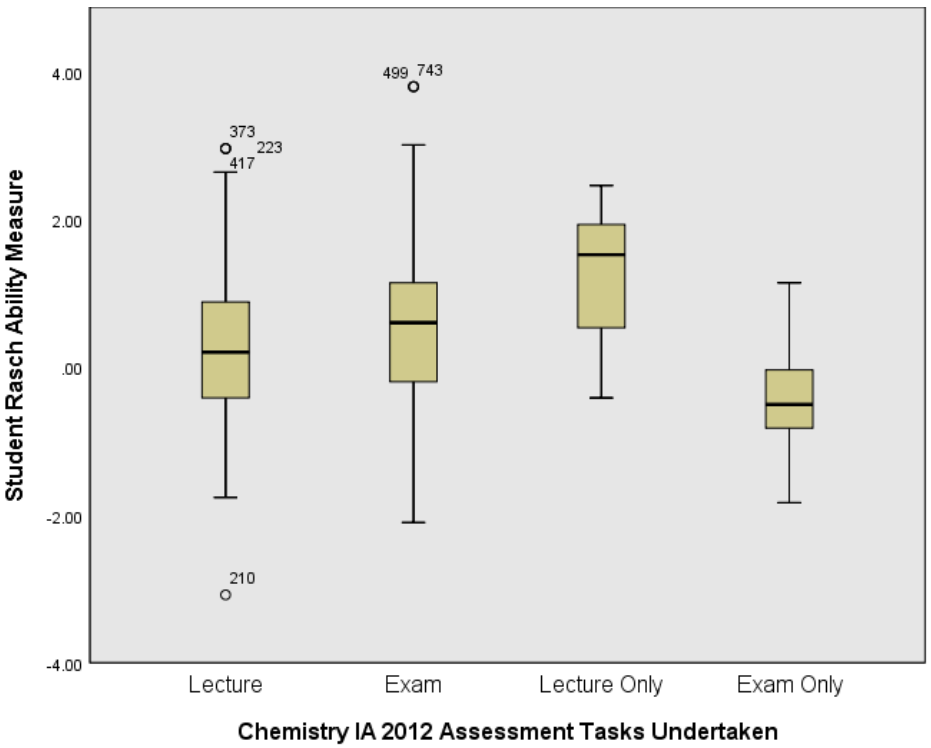

Figure 34: Boxplot Distribution of the Student Ability Measures in Common Assessments from Chemistry IA 2012, showing the Ability of the Students who undertook both Test and Retest and those who undertook one Assessment Task

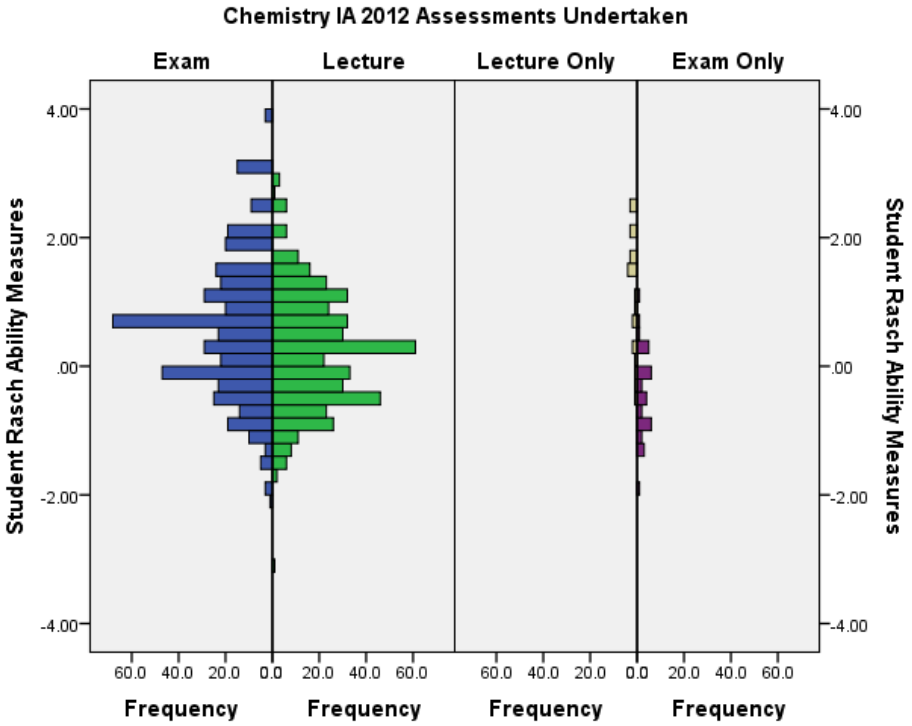

Figure 35: Histogram Distribution of the Student Ability Measures in Common Assessments from Chemistry IA 2012, showing the Ability of the Students who undertook both Test and Retest and those who undertook one Assessment Task

The differences between the student results in their first attempt during the lecture tests and their second attempt within the redeemable section of the final exam can clearly be seen within these figures. It illustrates that even though the range of the student abilities measured within the two tasks is still roughly the same, more of the students lie towards higher ability measures when they attempt the assessment task the second time. These figures also clearly show the disparity in performance between the students who undertake only one of the two assessment tasks. It shows that students who only undertook the redeemable section of the final exam had a lower average ability measure than any of the other student cohorts and students who only undertook the lecture tests were some of the highest ability students. It can also be seen that while the range of student ability measures between the lecture tests and the redeemable section of the final exam is similar, the student cohort is shifted toward higher ability measures within the redeemable section of the final exam which results in the statistically significant difference that is observed. These trends were also observed within the other assessment tasks being compared, which show the slight shift in the student ability between the lecture tests and the redeemable exam, as well as the large shift between the students who only undertook one assessment task (see Appendix 7.22 for the distribution plots of the student ability measures from all the assessments compared).

It needs to be remembered that there is the potential that not all of the students are improving between the two assessment tasks, and therefore it is possible that the statistically significant result observed is not representative of the experience of all the students. In the same way as was done for CTT, the numerical values associated with the changes in student performance, as well as values associated with the performance of the students who only undertook one assessment, can be generated, and compared. This provides information about the number of the students that are showing changes in their ability and the direction of that change, and how large those changes are. It

also gives further insight into the students who are only undertaking one of the assessments and how that influences their outcomes within the assessment. An example of this analysis can be seen in Table 33.

Table 33: Average Student Ability Measure from Shared Assessment Tasks within Chemistry IA from 2012 and How they Shift with Changes in Student Performance

		Count	Lecture Test Average Student Ability	Lecture Test S.D.	Exam Average Student Ability	Exam S.D.	Average Ability Change
Student Cohorts	**Cohort**	511	0.26	0.90	0.58	1.07	
	Lecture Only	24	1.16	0.96			
	Exam Only	34			-0.43	0.66	
Test-Retest Change	**Increase**	318	0.21	0.91	0.90	1.00	0.69
	Decrease	135	0.37	0.85	-0.17	0.85	-0.54

The average change in the student ability measures is very similar between the students who improve and the students whose ability degrades; however, the number of students who improve compared to the number that degrade is what causes the statistically significant difference between the mean ability measures of the two assessment tasks. It can also be observed that the highest average student ability is observed for the students who only undertake the lecture tests, and the lowest average student ability is found within the group of students who only sat the redeemable section of the final exam. These results are echoed throughout all of the assessment tasks that were compared within this research, as the average ability change remains relatively constant across all of the comparisons being made and, in each case, the average increase is matched closely by the average decrease in student ability (see Appendix 7.23 for the results from all other assessments analysed). The highest average student ability measure is always seen within the student cohort who only sat the lecture tests, whereas the lowest average student ability measure is always observed within the student cohort who only undertook the redeemable section of the final exam assessment task.

There are several factors that may boost the average result of the lecture test only cohort that should be considered. The students who only undertook the lecture tests would have had every opportunity to re-sit the assessment within the final exam. Despite this, these students chose not to attempt to improve their result, which is reasonable if a student believes that they would not be able to improve upon their original result, or if despite their potential to improve they choose to focus on other aspects of the final exam and settle for their previous result. If the students felt that they had any room for improvement in their lecture test results, whether that be because they didn't score full marks or if they thought that they simply underperformed, then it would be reasonable for them to at least attempt the redeemable section, as only their best result would be used. Thus, this means that the students who achieved high results within the lecture tests are the most likely to skip their opportunity to improve within the redeemable section of the exam, which results in some of the highest ability measures being seen in the lecture test only student cohort. The number of students who only undertook one of the two assessment tasks is relatively small when compared to the overall cohort sample size, which indicates that only a small fraction of the student cohort do not take advantage of the redeemable nature of the assessment. The small sample size also means that these averages are more easily influenced, and thus extremely high achieving students and extremely low achieving students present within those cohorts may shift the

average measure. The relative positions of the students who sat only one assessment to the rest of the student cohort can be somewhat seen within Figure 34; however, it is difficult to accurately compare the distribution of the students who only undertook the assessment task once to those whose who undertook both using that figure. A better methodology is seen within Figure 36 and Figure 37, which shows the lecture test and redeemable exam measures and more clearly demonstrates where the students who only undertake one assessment lie in comparison to the rest of the student cohort.

These figures clearly highlight that it does tend to be the higher ability students who make the choice not to undertake the redeemable section of the final exam, although there is a small number of students around the average ability level that seemed to make this choice also. Conversely, the students who only undertook the redeemable part of the exam can be seen to lie below the average student ability level. It should be noted that this is not true of all the students who only undertook the redeemable exam, but 32 of the 34 students who only undertook the redeemable section of the final exam had an ability measure lower than the mean ability measure of the student cohort who undertook both assessment tasks. Based on the previously discussed results, this trend is consistently observed in all of the comparisons made within this research (see Appendix 7.24 for the distribution comparisons of all the other assessments analysed), which suggests that the reasons for this occurrence are present within all of the courses being analysed and are not unique to this specific case.

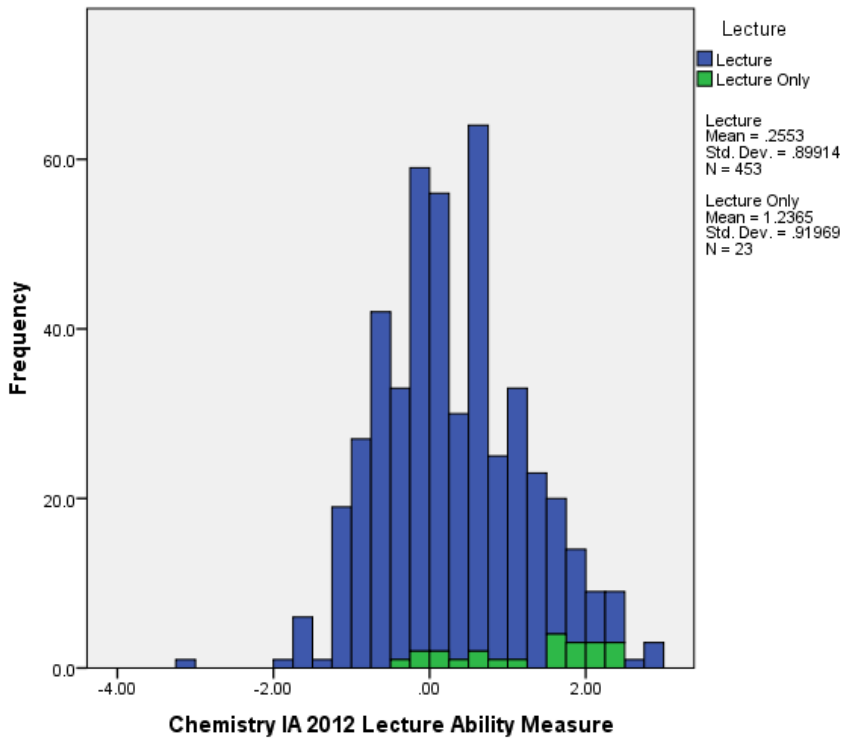

Figure 36: Ability Measure of the Students in Lecture Tests within Chemistry IA in 2012, Comparing the Ability of Students who only undertook the Lecture Test (Lecture Only) to those who undertook both Assessment Tasks (Lecture)

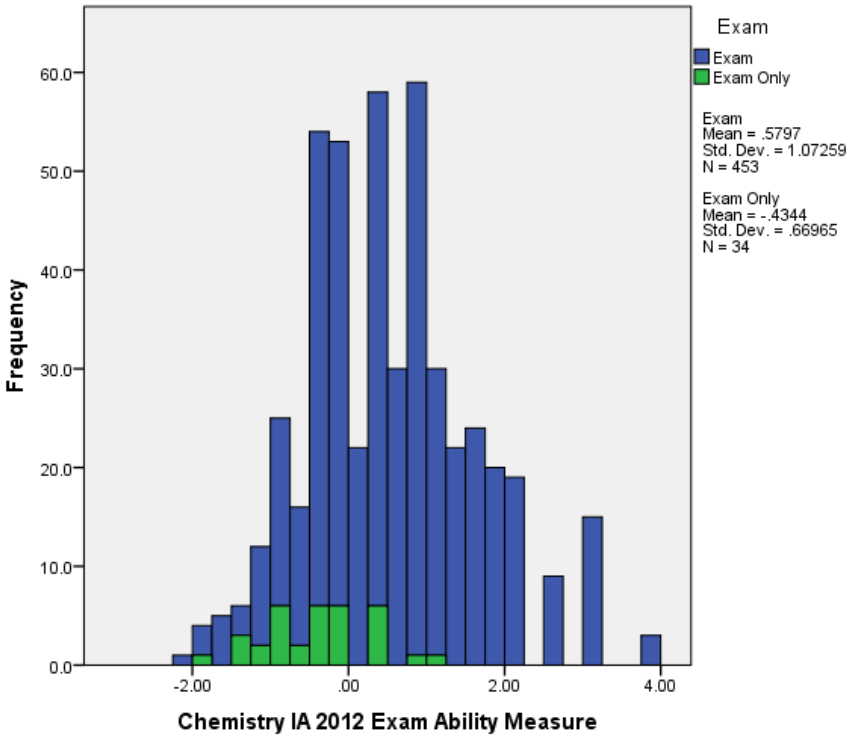

Figure 37: Ability Measure of the Students in the Redeemable Section of the Final Exam within Chemistry IA in 2012, Comparing the Ability of Students who only undertook the Lecture Test (Lecture Only) to those who undertook both Assessment Tasks (Lecture)

Whether the experience of undertaking a MCQ assessment is helping the students to improve upon their result within the redeemable part of the exam, or if the students simply perform better after all the content has been delivered, cannot be known through this research. What is clearly seen, and demonstrated by this comparison, is that given the opportunity to re-sit an assessment the majority of the students will take that opportunity and it is likely that there will be a positive statistically significant difference in the results of those students. The fact that not all the students see improvement should not be a concern as there are several variables around the mentality and the behaviour of the students that cannot be tracked based purely on their results. These changes in performance will vary from student to student and can be influenced by factors that are completely unrelated to the course itself, and thus there is no feasible way to account for them within this research or within the assessment tasks themselves. While the use of a test-retest format may cause concerns for the validity of assessments, there should be no doubt that this will positively influence the results of the students.

4.3 Comparison of Student Ability between Yearly Cohorts

4.3.1 Assumptions and Methodology

Comparing student ability between yearly cohorts follows a similar logic to comparing student ability within the same semester, as it is based on the ability to 'anchor' multiple assessment tasks together using shared items. Multiple items are reused over multiple years in chemistry assessments at The University of Adelaide as the core course content does not change and thus there is no requirement to adjust the assessment tasks (which is why the items are not published after the assessment tasks or within this research, so that the integrity of the items is maintained). It is possible to use these items to link the results of the students, allowing for a comparison of student cohorts over multiple

years. This could potentially give information about changes in the ability of the students enrolling within the course, and if there are any changes in the performance of the assessment items over time, and can provide evidence either for or against the anecdotal suggestion that students are performing less well each year. When comparing the results of one year to another all the MCQ assessment tasks undertaken in both years need to be included within the analysis, as this is the only way to account for potential differences in how the student cohorts approached the redeemable nature of the assessment tasks.

Before the student's ability can be compared there are several assumptions that need to be addressed either through analysis or by justifying their rationale. The largest assumption is that it is reasonable to compare student cohorts between years, as differences between years that are unrelated to the student cohort may result in differences being observed that are not reflective of the ability of the student cohort. This could relate to changes in how the course is structured, how it is taught, or who is teaching the students, as any of these factors has the potential influence the student ability determined through the use of assessment tasks. This assumption means that every student cohort has been taught the same content and their ability has been measured under the same conditions. Consistent measurement conditions are possible between years, as many items are shared between years and all the assessment tasks are undertaken under exam conditions. The consistency of the content being taught and how it is taught to the students is harder to justify; however, unless a course undergoes large changes (which would likely affect the items used within the assessment tasks) it is reasonable to assume that each year the same content is presented to the students in a similar manner. Another consideration that needs to be addressed is how much variance is expected to be seen between the yearly student cohorts, as this will inform the expectations of comparison between the student cohorts. The use of prerequisites can be employed to ensure that the students have previously displayed a certain level of competency in the subject, and thus this ensures a minimum ability level across the student cohort. As the courses being analysed within this research are first year chemistry courses it means that any prerequisites will be subjects from high school. Two of the courses being analysed (Chemistry IA and Chemistry IB) have year 12 chemistry with a specified minimum level of achievement as a prerequisite(even if what the students learn within the prerequisite changes), while the other two courses (Foundations of Chemistry IA and Foundations of Chemistry IB) require no prior chemistry experience. Based on this, it can be reasonably assumed that the students enrolling in Chemistry IA and IB are more likely to show a level of consistency in their ability across multiple years due to their prerequisite requirements. Conversely, it might be expected that there will be some amount of variation within Foundations of Chemistry IA and IB student cohorts due to the lack of prerequisites, which means that any student can enrol regardless of their previous experience with chemistry content (provided that they did not meet the entry requirements for Chemistry IA and Chemistry IB). While these are reasonable assumptions to make and justify, there is no way to determine if the teaching methodology, the student experience, or the student enrolment varies enough between years that it meaningfully influences the ability of the student cohort. What can be justified analytically is if the performance of the items is consistent from year to year by comparing the item difficulty. Confirming that the individual assessment items are performing in the same way within every student cohort ensures that it is the students themselves causing any changes seen and not the items. Confirming that the items are not changing requires the use of a paired-sample t-test, as this can be used to determine if the performance of an item changes significantly between years. As outlined above (see Section 4.2.1) CTT can only compare items that are shared between assessment tasks, while Rasch analysis can compare the entire assessment task due to it generating independent measures of item difficulty and student ability.

4.3.2 Classical Test Theory

The first step in comparing the ability of the students is to ensure that the behaviour of the items is consistent between years to ensure that they are not influencing the comparison of student ability between years. The first step in this comparison is determining the number of items that are shared across the years being compared, as CTT can only compare results that are generated using the same items. The summary of the number of items shared between all the years being compared within this research can be seen in Table 34, where the maximum number of items that could be shared across all of the years was 90.

Table 34: The Number of Items that are Shared Between MCQ Assessment Tasks Undertaken in Different Years (2012-2015) within each Course being Analysed

	Number of Shared Items Across All Years
Chemistry IA	52
Chemistry IB	67
Foundations of Chemistry IA	63
Foundations of Chemistry IB	47

It is expected that the mean item difficulty does not change between years (which is the null hypothesis within the comparison being made) as there is no reason for the items' behaviour to change between years. The issue with comparing the behaviour of the mean item difficulty generated by CTT is that the item difficulty is influenced by the student cohort, and thus when comparing the mean item difficulty it is possible that any differences observed are the result of changes in student performance rather than any change in the item behaviour. Therefore, any statistically significant difference seen within the mean item difficulty may be an indicator of statistically significant changes in student cohort ability and thus should be rationalised using both the results of the item difficulty comparison and the student ability comparison.

Despite these concerns with comparing CTT mean item difficulty values over multiple years, it is important that the comparison is made to ensure that any statistically significant differences are highlighted and are known when the student cohort abilities are compared. The CTT mean item difficulty comparison, undertaken using a paired-sample t-test of all the shared items used within MCQ assessments across all the years analysed within this research (2012 – 2015) can be seen in Table 35.

Comparing the mean item difficulty shows that there is statistically significant differences in the behaviour of the items between years; however, as mentioned previously, the fact that the item difficulty values are dependent upon the student cohort means that the statistically significant differences observed may be an indicator of a change in student performance rather than a change in item difficulty. If the difference in item difficulty is the result of consistent increase or decrease across the entire student cohort, then the item difficulty is likely to change consistently across all of the items to reflect that difference (e.g. if the student cohort has a significantly higher ability one year it should be expected that all of the item difficulties will be reported as higher values [higher values means more students answered correctly]); however, if the statistical significance observed within the comparison of the mean item difficulties is the result of particular items showing large changes between years then it is unlikely that this is reflective of a change across the entire student cohort. Whether all the item difficulties are shifting or only a select few items change between years can be observed using a scatter plot that compares the individual item difficulty values of shared

items across the years being compared, as this will highlight any individual items that show large differences. Using the Chemistry IA results as an example of this the item difficulty of shared items from 2012 MCQ assessment tasks (x-axis) is compared against the item difficulty of shared items from other years (2013, 2014, and 2015 shared item difficulty shown on the y-axis), as shown in Figure 38.

Table 35: Comparison of the Mean Item Difficulty Generated Using Classical Test Theory between all the Shared Items used within MCQ Assessment Tasks in each Year. Highlighted Cells indicate Observation of a Statistically Significant Difference

		2013		2014		2015	
		df	***p*-value**	**df**	***p*-value**	**df**	***p*-value**
Chem IA	**2012**	51	0.474	51	0.284	51	0.346
	2013			51	0.539	51	0.649
	2014					51	0.749
Chem IB	**2012**	66	<<0.001	66	0.001	66	0.036
	2013			66	0.072	66	0.032
	2014					66	0.351
FoC IA	**2012**	62	<<0.001	62	<<0.001	62	<<0.001
	2013			62	0.428	62	0.420
	2014					62	0.992
FoC IB	**2012**	48	0.769	48	0.368	48	0.062
	2013			48	0.048	48	0.001
	2014					48	0.067

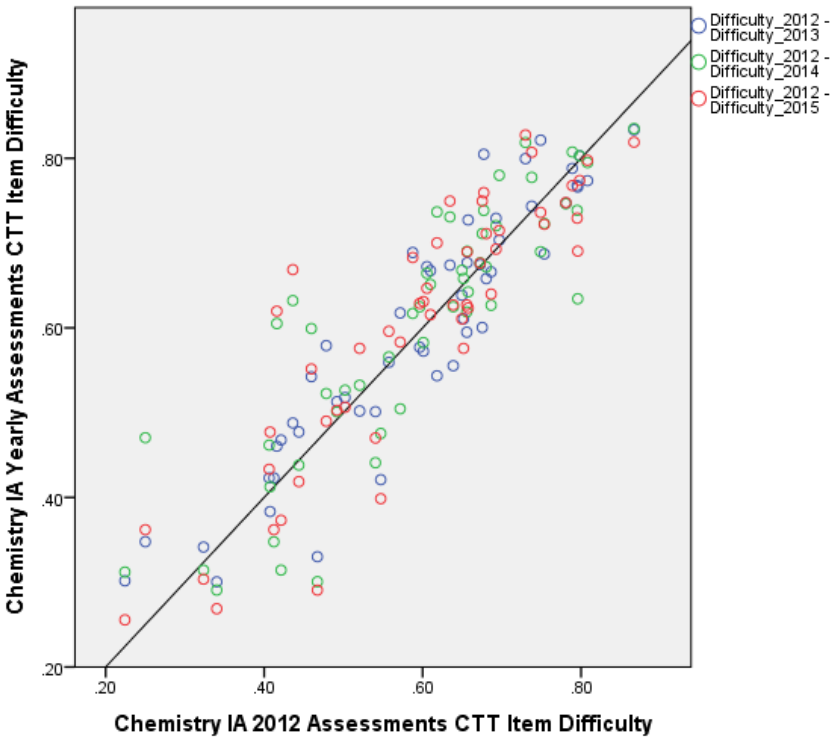

Figure 38: Scatter Plot of Shared Item Difficulty Values from Chemistry IA MCQ Assessment Tasks Undertaken in the Years being Compared (2012 -2015), Where the Item Difficulty Values are being Compared to the Values Generated from 2012 Assessment Tasks

The figure shows there is scatter around both sides of the line that indicates that items are performing the same within both assessment tasks, which in turn indicates that there is not a consistent shift within the item difficulties between years. While it is not expected that a consistent shift is observed, if this was observed it may indicate a difference between the student cohorts in the two years being compared. However, as none of the assessment tasks analysed within this figure showed a statistically significant difference between the mean item difficulty, the variance around the line is expected as it means that any differences in behaviour in one direction are matched by a difference in the other direction. Viewing the corresponding scatter plots for the other courses being analysed (see Appendix 7.25) reveals a similar trend, in that there is a large amount of scatter around the line representing the same behaviour in both assessment tasks too. It should be mentioned that just because the item difficulty changes are somewhat sporadic, it does not mean that changes in student performance are not influencing the item difficulty, as student performance does not have to increase in every facet of the course and may show fluctuations across different topics. That being the case, it cannot be known exactly what is causing the changes in item difficulty, but what is known is that within some of the courses there is a statistically significant difference in the mean item difficulty between years. While this has the potential to cause problems within the student analysis there is no way to account for the years that have mean item difficulties that are statistically significantly different, and as it seems more likely that these differences are caused by changes in the student cohort rather than changes in the item behaviour these results can only be interpreted once the student cohorts are compared.

After comparing the behaviour of the items between years using the mean item difficulty, the student cohorts need to be compared; however, the redeemable nature of the assessments has the potential to cause problems as the cumulative raw score of the students cannot be compared as not all of the students will have undertaken the same number of items. This is due to some students not undertaking either the lecture tests or the redeemable exam section, which means that the total number of items that those students answered will be less than the students who sat all of the MCQ assessments, and thus if raw scores were compared those students would be expected to show significantly lower results. One way to compensate for this is to calculate the student's percentage score based on all the items shared between the assessment tasks they answered, as this means that regardless of how many items the students answered it provides a comparable measure of their performance within the assessment tasks. The other possible way that this issue could be resolved is to compare each assessment task individually; however, as mentioned earlier, this introduces the potential for student approaches to the assessments to influence the student outcomes and thus this method was not used. The comparison of the mean student percentage scores, undertaken using a paired-sample t-test, on items that are shared across multiple years can be seen in Table 36, where it was expected that the student cohorts will perform the same each year despite their differences.

Table 36: Comparison of the Mean Student Percentage Results on Shared MCQ Items used over Multiple Years within First-Year Chemistry Courses at The University of Adelaide. Highlighted Cells indicate Observation of a Statistically Significant Difference

		Mean Student Percentage	2013		2014		2015	
			d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value
Chemistry IA	2012	58.9	1035	0.586	1043	0.221	1076	0.084
	2013	59.5			1044	0.516	1077	0.262
	2014	60.2					1085	0.644
	2015	60.7						
Chemistry IB	2012	54.9	896	0.009	933	0.003	936	0.143
	2013	57.8			941	0.778	944	0.233
	2014	58.2					981	0.132
	2015	56.5						
Foundations of Chemistry IA	2012	62.3	700	0.634	663	0.065	708	0.412
	2013	61.7			719	0.139	764	0.703
	2014	59.9					727	0.281
	2015	61.3						
Foundations of Chemistry IB	2012	61.8	597	0.069	565	0.618	588	0.763
	2013	64.5			600	0.186	623	0.114
	2014	62.6					591	0.830
	2015	62.2						

Despite the statistically significant differences observed within the results of the mean item difficulty comparison (where 9 statistically significant results were identified) there is only 1 student cohort that shows significant differences from other student cohorts. Even though there was the potential for more variations to be observed with the ability of the student cohort from Foundations courses (due to their lack of prerequisites allowing for more variety within the educational background of the students who enrol in those courses), neither of those courses showed any statistically significant difference in the percentage scores of the student cohorts compared. The only two statistically significantly different performances are the comparisons of the student cohorts from Chemistry IB (2012 with 2013 and 2012 with 2014). Given that the performance of the Chemistry IB 2013 student cohort is not statistically significantly different to that of the 2014 student cohort this suggests that the only real outlier is the 2012 student cohort within Chemistry IB. It can be seen using the average percentage results of the students that the 2012 cohort performed less well than the other cohorts analysed. The fact that 2012 is the only year that shows significant deviations in student results also suggests that the significant differences observed within the item difficulties are a result of the items and not the students. Despite the lack of statistically significant differences observed within the other years it is possible that there are variations in the student cohort that may still influence the mean item difficulty enough to cause statistically significance differences to be observed within the comparison of the mean item difficulty. However, if this is not the case then the behaviour of the items must be shifting in some way between years to cause the statistically significant difference observed within the mean item difficulty. If the behaviour of the item is changing between years, it means that the comparison of the student cohorts is not valid as they are

not being assessed in the same way each year. There is no further analysis that can be undertaken using CTT to confirm or refute the influence of the item behaviour on the student cohorts, and thus it has to be assumed that for the purposes of this analysis that it is the student cohorts that cause the difference. This assumption can never be completely justified using CTT, and thus does place some doubt on the rest of the results seen within this comparison.

It is unknown whether there were any significant disruptions or changes to the Chemistry IB course in 2012 that influenced the results of the students causing the statistically significant differences observed; however, it seems unlikely that this is the reason for the statistically significant difference as there are no differences between the 2012 and 2015 student cohorts. This suggests that either the statistically significant difference observed was due to a difference in the ability of the students who enrolled, or potentially the influential factor was within 2013 and 2014 resulting in statistically significantly higher abilities within the student cohort in those years. To gain a better understanding of the possible origin of any differences observed, a plot of the distribution of student performances can be used to compare the results of the students. This has the potential to inform whether the difference is a result of a shift within the entire student cohort, or within a specific group of students. The student distribution comparison for Chemistry IB can be seen in Figure 39 and Figure 40.

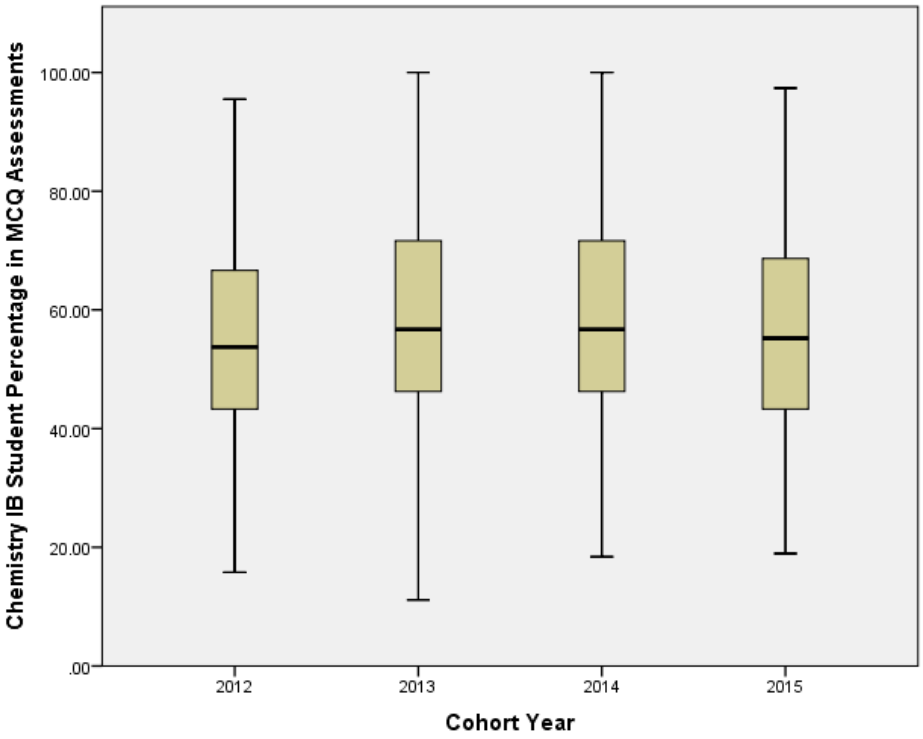

Figure 39: Boxplot Distribution of Student Percentage Score on Shared MCQ Assessment Items Undertaken within Chemistry IB

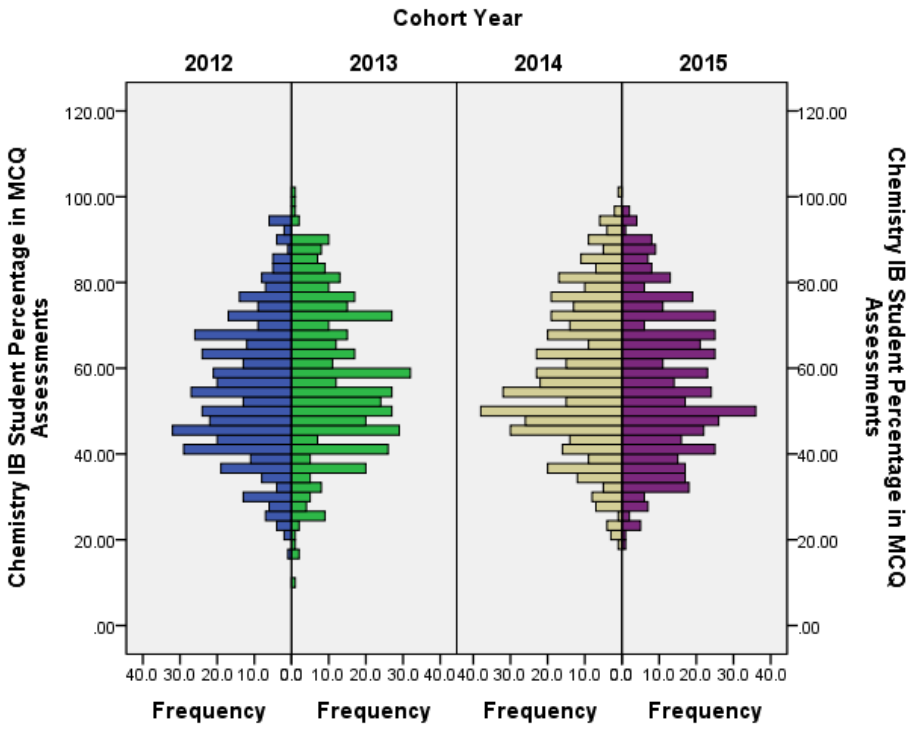

Figure 40: Histogram Distribution of Student Percentage Score on Shared MCQ Assessment Items Undertaken within Chemistry IB

Based on these comparisons of the Chemistry IB student performance distributions, it appears as though the major difference within the 2012 cohort compared to the other years is that is has fewer high ability students. This causes a lower average percentage score overall, which then contributes to the significant differences observed. It is important to remember that even though the mean difference between the student cohort percentages is quite small (a difference of approximately 3% showed significance) the number of students being compared is very large, and thus it is expected that small deviations will result in identification of statistical significance. While usually this is a cause for concern within statistical comparisons as it results in almost every comparison showing significance due to natural variation, in this instance these issues were not observed. This can be justified by the fact that only one cohort showed significant differences from the other cohorts despite the large number of cohorts being compared. In addition to this, the cohort that was significantly different can be visually observed to show deviation away from the other student cohorts analysed (see Appendix 7.26 for the student distribution from the other courses analysed). To determine what makes that cohort perform differently from the other cohorts would require a deeper analysis into the students that enrolled into the course and potential factors that may have differed within that year compared to the other years, which is beyond the scope of this research. What this research does inform is that despite anecdotal suggestions that students tend to perform less well every year, this is not the case, and it can be confirmed that (at least over the four-year period studied here) that this shift is not a consistent trend downward that is not large enough to result in statistical significance when yearly cohorts are compared.

4.3.3 Rasch Analysis

Comparison of the performance of yearly cohorts was also undertaken using the Rasch model in which, due to its nature, there are differences in how the assessment tasks, items, and the student cohort can be treated. To start with, the item difficulty measures of the shared items need to be

compared to ensure that they are performing the same way in all the assessment tasks to validate them being used to link the assessments. To obtain comparable item difficulty measures the assessments need be linked, which is done using items that are common to all the assessments. Within this research the assessments were linked by stacking the shared items within the analysis, which meant that the shared items were treated as if every student cohort had undertaken them within the same assessment task. Using labels included within the stacked dataset it is possible to generate item and student measures that are purely based upon the results from one specific year that still lie within the same scale as the rest of the dataset. If this method is used to generate measures for each year being compared, it provides measures unique to each year that still lie within the same scale as the other years, which means that they are directly comparable to one another. A benefit of using the Rasch model is that due to the independence of student ability and item difficulty measures, there is no requirement to only include items that are shared across all the assessment tasks. It also means that unlike CTT, where the percentage score had to be used to compare the results of the students due to the potential differences in the number of items answered, the student ability measures are comparable regardless of differences in the number of items answered. Even though the entire assessment can be used in the comparison, only the items that are shared between assessment tasks can be compared, as the other items are not comparable measures. Due to the independence of the student ability measures it is not important if the assessment items themselves are significantly different as long as the items that are shared between the assessment tasks can be used to link the assessments. The comparison of the mean item difficulty of the shared items was done using a paired sample t-test and can be seen in Table 37, where it is expected that the mean item difficulty of the shared items does not change between years. The stacked analysis item difficulty, which represents the item difficulty based on all four years analysed, was also included within the comparison to determine if any of the individual year results had a significant influence on the linking of the assessments, as this could potentially cause errors in the significance testing.

From the data presented it is clear that no statistically significant differences are observed in the comparison of the mean item difficulty of the shared items when comparing the Rasch item difficulty measures produced each year. The stacked analysis item difficulty was also included within the comparison to determine if any of the years had a significant influence on the linking of the assessments, as this could potentially cause errors in the significance testing. The fact that none of the item difficulties showed any significant differences between years is a different result to what was obtained using CTT analysis; however, the independence of the Rasch item difficulty measures is likely to be responsible for this outcome. This is because it is highly likely that the results of the item difficulty comparison made using CTT were influenced by the student cohort; however, it is impossible for the student cohort to influence the results of the item difficulty comparison when Rasch analysis is used due to the independence of the measures.

Table 37: Comparison of the Mean Item Difficulty Measures Generated using Rasch Analysis Comparing the Items that are Shared Across Yearly Assessment Tasks. Highlighted Cells indicate Observation of a Statistically Significant Difference

		2012		2013		2014		2015	
		d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Chemistry IA	**Stacked**	69	0.931	69	0.843	69	0.989	69	0.941
	2012			68	0.767	53	0.880	51	0.989
	2013					54	0.973	52	0.896
	2014							67	0.872
Chemistry IB	**Stacked**	68	0.757	69	0.927	69	0.864	69	0.978
	2012			68	0.693	68	0.923	66	0.889
	2013					69	0.857	67	0.973
	2014							67	0.974
Foundations of Chemistry IA	**Stacked**	69	0.464	69	0.785	69	0.792	69	0.902
	2012			62	0.530	62	0.656	62	0.620
	2013					69	0.757	69	0.944
	2014							69	0.805
Foundations of Chemistry IB	**Stacked**	68	0.951	67	0.753	69	0.498	69	0.639
	2012			53	0.745	48	0.525	48	0.291
	2013					56	0.368	56	0.491
	2014							69	0.481

Another way of comparing the item difficulties using Rasch analysis is with a differential item functioning (DIF) plot, which does not identify significant differences between items but does visually show items that are performing distinctly differently from year to year. This method was not used extensively in this research due to there being no significance observed between the yearly item difficulties. Depending on the requirements of the analysis, a DIF plot may be a more effective and easier method of comparing the items, as in many cases any significant differences between two items' difficulties (a deviation ≥ 0.50 logits) can be easily seen within the plot. That being said, when more items and assessments are included within the comparison, it becomes harder to determine differences based on the plot alone (another reason that the DIF plot was only used as a quick reference point within this research), as the plot becomes congested and there is no longer simply two points to be compared but rather one point for each assessment being analysed. This can be seen in Figure 41, which is the DIF plot for all the shared items within Chemistry IA MCQ assessments, including four years' worth of assessments. As a result, no effective information can be obtained from the plot except that all of the items lie closely within the same item difficulty range, meaning there are no obvious outliers to be analysed further (as noted previously there is no reason for the points to be connected; however, this does make the graph easier to follow and thus this was included).

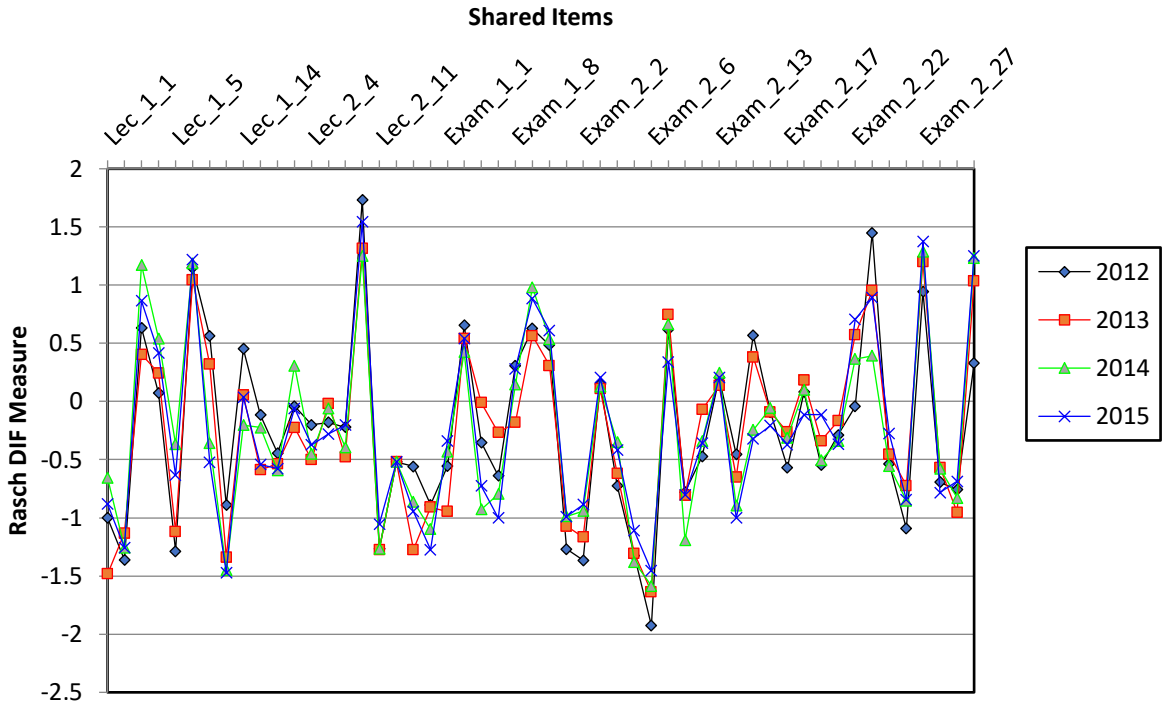

Figure 41: DIF Plot of the Shared Items Asked in Chemistry IA MCQ Assessments in all Four Years Analysed to Visually View the Differences in Item Difficulty

Viewing the plot it can be seen that despite there being minor fluctuations within the Rasch item difficulty measures, none of the fluctuations appear to surpass the 0.50 difference in DIF measure required to constitute a statistically significant difference. A DIF plot could be used either before a statistical comparison, to highlight items that are obviously deviating between assessments, or it may be used in conjunction with the statistical analysis. Another reason that the DIF plot was not used extensively within this research is that a scatterplot of the item difficulties was used as an alternative method to ensure that the individual items were not significantly changing. The scatterplot generated matches what was produced when comparing the individual assessment items using CTT, where the comparison of the Rasch item difficulty measures from Chemistry IA shared items can be seen in Figure 42.

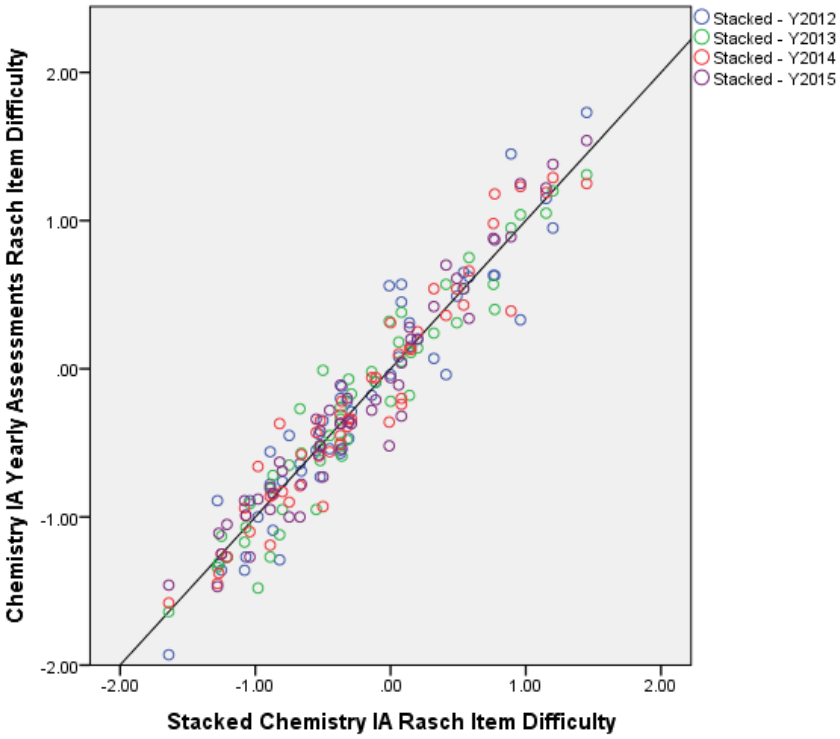

Figure 42: Rasch Item Difficulty Measure Scatterplot of Shared MCQ Assessment Items from Chemistry IA using all Years Analysed (2012-2015) Compared Against the Rasch Item Difficulty Measure Produced by the Results from Each Individual Year

While there is some deviation within the individual item difficulty measures, none of them show a large amount of deviation from the line that shows when the item difficulty measures from each year match the measure generated from the stacked dataset. This finding is matched by the cross-plots of the other courses analysed (see Appendix 7.27). Based on these results, the Rasch item difficulty measure of individual items is not changing between years, and thus can be used to link the assessment tasks between the years without having any influence on other measures. Knowing that the items are not influencing the linking of the student cohorts means that the student ability measures can be compared across all the years being analysed without any concern for unrelated influences. The student ability measures generated from the assessment tasks do not need to be adjusted in any way to account for any differences in the items used in each assessment task due to the independence of the student ability measures from the item difficulty measures. Thus, as long as the student ability measures are placed on the same scale by linking the assessment tasks using shared items, they can be compared over multiple years despite the same items not appearing in every year. The comparison of the mean student ability of the student cohorts between 2012 – 2015 covering all the courses analysed can be done using an independent sample t-test that assumes that the mean ability of the student cohorts is the same every year being compared, the results of which can be seen in Table 38.

Table 38: Comparison of the Mean Student Ability Measures from Yearly Student Cohorts on MCQ Assessments over Multiple Years within First-Year Chemistry Courses at The University of Adelaide. Highlighted Cells indicate Observation of a Statistically Significant Difference

		Student Cohort Mean Ability	2013		2014		2015	
			d.f.	*p*-value	d.f.	*p*-value	d.f.	*p*-value
Chemistry IA	2012	0.210	1036	0.73	1044	0.579	1077	0.659
	2013	0.190			1044	0.366	1077	0.417
	2014	0.240					1085	0.884
	2015	0.233						
Chemistry IB	2012	0.302	896	0.013	933	0.002	936	0.153
	2013	0.451			941	0.565	944	0.279
	2014	0.486					981	0.093
	2015	0.386						
Foundations of Chemistry IA	2012	0.756	700	0.149	663	0.133	709	0.284
	2013	0.643			719	0.908	765	0.723
	2014	0.635					728	0.650
	2015	0.670						
Foundations of Chemistry IB	2012	0.634	597	0.032	565	0.016	588	0.012
	2013	0.438			600	0.640	623	0.669
	2014	0.397					591	0.946
	2015	0.403						

The table shows that there are two student cohorts that performed statistically significantly differently from other years; however, none of the other cohorts compared showed a statistically significant difference. The same difference in the 2012 Chemistry IB student cohort that was observed in the CTT analysis was observed in this comparison; however, it also shows a significant difference in the 2012 student cohort within Foundations of Chemistry IB from all the other years included within the analysis. The Chemistry IB 2012 cohort was seen to have a lower average Rasch student ability measure than that obtained by the 2013 and 2014 student cohorts, while the 2012 Foundations of Chemistry IB student cohort had a higher average Rasch student ability measure than the other cohorts analysed. The consistency across all of the years outside of these identified differences (only 5 out 24 comparisons showed statistically significant differences in the mean Rasch ability measures of the student cohort) shows that the average ability of the student cohorts closely match across the years being analysed; however, it is important to consider if the distribution of the abilities also closely match. This can be done by viewing the distribution of the Rasch student ability measures within each cohort using either a boxplot or a histogram. Both plots for the Foundations of Chemistry IB student cohorts are shown in Figure 43 and Figure 44.

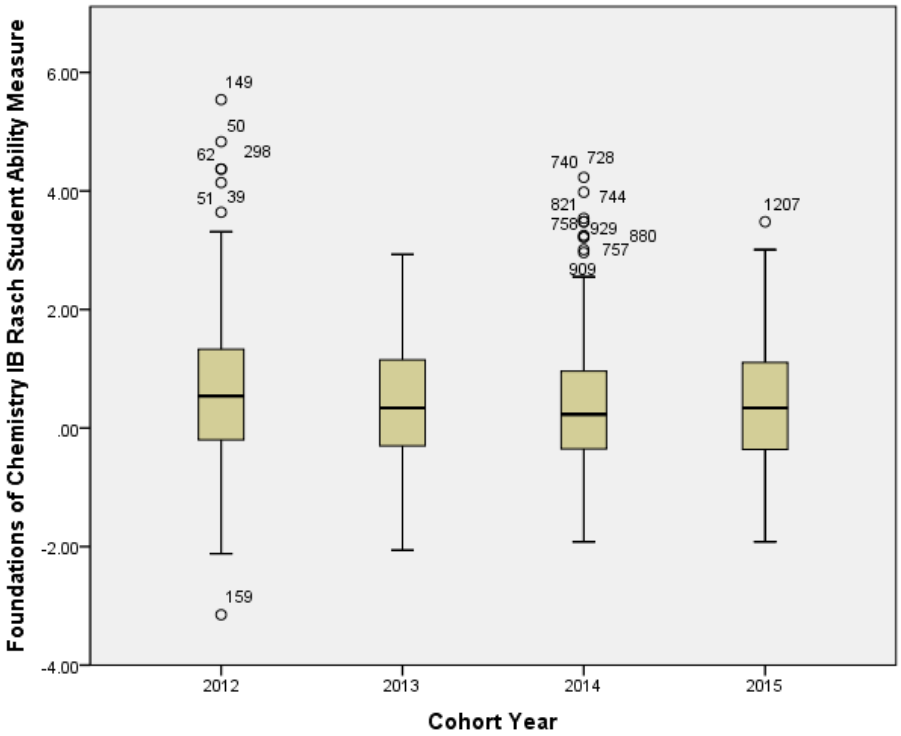

Figure 43: Boxplot Comparison of the Foundations of Chemistry IB Rasch Student Ability Measures from MCQ Assessments Between 2012 – 2015

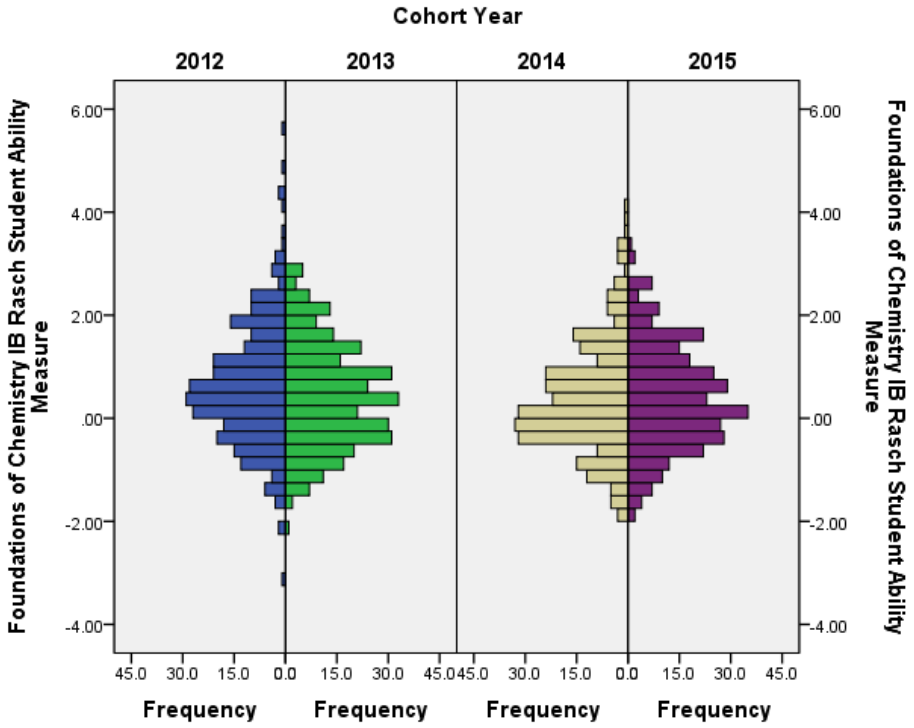

Figure 44: Histogram Distribution Comparison of the Foundations of Chemistry IB Rasch Student Ability Measures from MCQ Assessments Between 2012 – 2015

There is a slight shift in the Rasch student ability measure in Foundations of Chemistry IB during 2012 that meant that more students had higher Rasch ability measures and this resulted in its statistically significant difference from the other student cohorts, which is best illustrated within the

boxplot comparisons of the student cohorts but can also be observed within the histogram distribution. The 2012 student cohort only contains one outlier that was below the minimum whisker within the boxplot; however, the higher range covered within the 2012 whiskers in the boxplot would include several outliers seen within other years. This indicates that the number of students who obtained higher ability measures was greater within the 2012 student cohort than the other years analysed, as these higher ability students were not treated as outliers within the data, which resulted in the significant difference observed. The distribution of the student cohorts within other years closely match each other, following an approximately normal distribution, which illustrates why these student cohorts were not identified as being statistically significantly different from each other. It cannot be known what caused the statistically significant differences within the 2012 Chemistry IB and 2012 Foundations of Chemistry IB student cohorts based on this analysis alone (see Appendix 7.28 for the student cohorts from the other courses analysed); instead, this requires consideration of all the influences and changes that were specific to that year. It is possible, but unlikely, that the reason for the difference observed within the 2012 courses can be determined at this stage due to the amount of time that has passed. This highlights the importance of undertaking the analysis upon the completion of the assessment, as it is easier to identify the potential reasons for any differences observed immediately after completion. If a statistically significant difference is observed, then it may be important to attempt to identify what changes may have been made to the courses that have the potential to influence the results of the students. These changes may be accidentally influencing the students in ways that were not expected or intended, or they may be performing as they were expected to and resulted in a change within the student cohort. For example, the reason that the Foundations of Chemistry IB 2012 student cohort performed better than the other years analysed may be the result of a change implemented within 2013 and was carried through successive years that lowered the performance of those student cohorts. Depending on the change it may be considered reasonable that there is a slight shift downward in student performance; however, it is also possible that the change was aimed to help improve student performance and thus evaluating how the change affects the students is important when deciding whether it was a success. The analysis conducted here will not be able to identify the change to the course that resulted in the difference being observed, or even if the change was due to differences to the course or the student cohort, but it does inform whether a change occurred.

The results of Rasch analysis match closely with the CTT results, which show that there is no evidence to support the idea that student ability is decreasing each year. Of the two years that did show a statistically significant difference only one of them showed a decrease in the average student ability in the following years; the other saw an increase in the average student ability in future years. The number of differences observed is small compared to the number of comparisons made, and thus most of the evidence shows that student cohorts perform to the same level each year with a high degree of consistency. Therefore, it should be expected that, assuming there are no significant changes to the course material, the student cohorts will show similar performance outcomes each year.

4.3.4 Changes in Student Performance

The overall consistency in how the students perform between years should be expected given that there are generally no substantial changes in the courses themselves. Thus, if any differences are observed between two student cohorts this likely represents a difference in the educational background of the students enrolling in the course. There are two key factors that are largely responsible for the students having similar educational backgrounds; the first is the high number of students that enrol into first-year university courses either immediately after or within a couple

years of completing their high school education. While the students will not necessarily have undertaken the same subjects in high school, there are requirements that the students must meet to obtain certification of completion, such as competency in English and Mathematics subjects. The second reason is the use of prerequisites, as it means that the students have already shown a minimum competency within areas that are deemed relevant to the course. Despite the logic that prerequisites will provide a more consistent student ability within the enrolled student cohorts, a statistically significant difference was observed in courses that both did and did not have any prerequisites, and thus suggests that prerequisites will not always result in student cohorts of similar abilities. The variation observed in Foundations of Chemistry IB showed an increase in the student ability, which does not invalidate the assumption that the students have a minimum competency. Rather, there is potential that the 2012 student cohort, which exhibited a significantly higher ability, simply surpassed that minimum competency. The statistically significant decrease in the mean Rasch student ability measures observed within the 2012 Chemistry IB student cohort does seem to go against the idea of a base ability level due to the course prerequisites. However, upon reflection of the student cohort distribution, it does appear that the cause of the identified difference may be due to more higher-ability students present within 2012 compared to the other years analysed. Thus, the difference in ability level may not be due to changes in the mean ability of the student cohort, but rather it may be due to the fact that there were more higher ability students within 2012 that increased the mean ability of the 2012 student cohort to be statistically significantly higher than it was in other years. Without doing a deeper analysis into the students enrolled and any potential influences that were different between years, the exact reasoning for the difference cannot be known. What can be observed is that while there are deviations within the student cohorts between years, in the majority of cases these deviations are not enough to cause statistically significant differences in the performances of the student cohorts year to year.

An important consideration is that this analysis does not show the ability of the students across all the assessments that they undertook within the course and is only reflective of the student ability within MCQ assessment tasks. The reason that MCQ assessments are the best suited for this comparison is due to their objective marking and the consistency of their use across multiple years. The correct answer on a MCQ item will not change between years, and while the technically correct answer to written response style items also does not change between years it is possible that the person responsible for marking those items may change or the criteria they use may be slightly different between years. Any changes to how the students are marked between years would influence the ability measures obtained by the students for each assessment for reasons that are completely unrelated to their ability. Thus, while it is possible to use written response items to compare the ability of the students across multiple years, the possibility of subjective marking needs to be carefully considered before the comparison is undertaken, and in an ideal scenario the assessments from all the years being analysed would be remarked by the same person before the comparison is made. This would ensure that the marker and the marking criteria are consistent across all of the years being compared, and thus any changes to the student ability would be a result of changes within the student cohort and not changes to how the assessments were marked. This would obviously be an extremely lengthy undertaking, and one that is unnecessary when there are no concerns with the marking consistency of MCQ assessments due to their objective nature. While this does not alleviate the concern that there may be differences in the student cohort based on their results within other assessment formats, it highlights that using MCQ assessments is the most logical way of comparing the student across multiple years.

Knowing that there is often no statistically significant difference in the ability of student cohorts between years does show that there is a level of consistency within the relevant courses and how these are delivered to the students. Thus, in a way the observed overall consistency of the student cohort is reflective of the course and how changes to the course influence student outcomes. This means that there is the potential to use this comparison as a way of determining how differences in the course influence the students, and if those changes had their intended effect. These could include changes such as the weighting of assessments, presentation of lectures, tutorial assessments, or expectations of the student. Of course, not all the changes made to a course would be reflected within the results of MCQ assessment tasks, and this is therefore something that needs to be considered before this comparison is used to rationalise any outcomes. Being able to show that there is consistency in the student ability across multiple years is important because it shows that there is little change in students' abilities over time, and also because it shows that the course is able to provide consistent learning outcomes for the students.

4.4 Comparison of Student Performance in Different Courses

4.4.1 Assumptions when Comparing between Courses

In theory, it is possible to validly compare the performance of students from courses in different disciplinary areas if there is some way that the two courses can be linked together. This would enable a comparison of the student performance across the different courses and a comparison of the difficulty of assessment items across both courses. Many students that are enrolled in first year Chemistry courses at The University of Adelaide are also enrolled in first year Biology courses. The first-year Biology courses also utilise MCQ assessment tasks throughout the semester as a way of assessing the students; however, unlike the chemistry courses, these assessments are not redeemable in any way. The students common to both courses could be used to link the two courses together allowing for a comparison to be made between the results of the MCQ assessment tasks used in both courses. When comparing the courses, and particularly when linking using the results of the students, it is important to consider what the ability that is being measured relates to. This is a consideration of dimensionality and whether it is reasonable to assume that the results of the two courses give comparable outcomes. Some courses make more sense to compare than others: for example, it would be expected that the results of students in a language course do not measure the same latent ability trait that the results of a more physically involved course, such as metal work, do. That does not mean that there would be absolutely no ability overlap, as commonly it is thought anecdotally that high performing students in one course are also high performing students in other courses. However, it cannot be assumed that this expectation is solely a result of the student's ability; rather it may relate to their work ethic or some other factor that influences their performance within assessments. Courses that are much closer in their content and concepts would be expected to share a greater overlap in the ability that they measure. Thus, the comparison of Chemistry and Biology as two science-based courses is one of the more reasonable comparisons that could be made when comparing the results of assessment tasks between courses.

If the raw scores are used for the comparison of the students, it assumes that the items being asked within the assessment tasks are of equal, or at least comparable, difficulty, which cannot be confirmed without the presence of any shared items in the assessment tasks from the separate courses. As there are no shared assessment items between Chemistry and Biology courses there is no validity in a direct comparison of the difficulty of the assessment items, and thus it is unreasonable to compare the raw scores of the students expecting them to be comparable

measures. This means that linking the assessment tasks through Rasch analysis, using the students undertaking both courses, is required if comparable measures are hoped to be generated. If Rasch modelling is used, the dimensionality of the dataset that includes two difference courses has to be deeply considered, as one of the key underlying assumptions of the Rasch model is that student ability and item difficulty are unidimensional measures that can be expressed on the same scale. For this to be true, the student ability being measured within Chemistry assessments needs to be the same as the ability that is being measured within Biology assessments, which in turn means that the ability being measured needs to be representative of the student's ability within science or a representation of their academic ability. Whether this is a reasonable assumption or not would ideally be determined by linking the assessments using items and then using the dimensionality report from the analysis to determine the number of influential factors. However, as there are no items that are common between the courses and the anchoring needs to be conducted using the students, the comparison cannot be made in this way. Thus, for the comparison to be made it must be assumed that the ability trait being measured in both assessment tasks (i.e. that the assessment tasks from both Chemistry and Biology require the students to utilise similar abilities even if the content being assessed is different) is the same without being able to test that assumption in any way until after the analysis is conducted. If that assumption can be justified or is assumed to be true, the next step is to identify which students can be used to anchor the courses together. If the students used to anchor the assessment results have different ability levels between the two courses it would result in a skewed scale that will give results that are not reflective of where the student ability and item difficulty lie relative to each other in reality. Thus, the most important step when comparing the courses is identifying which students to use as anchors, as this will influence the scale; hence it is important that there is no bias introduced to the comparison due to the students used.

4.4.2 Issues Identified with this Comparison

The first approach taken within this research was to link the assessment tasks by anchoring specific student ability measures taken from the first assessment task into the second assessment task. This would then cause the logit scale within the second assessment task to shift such that the student ability and item difficulty measures are now placed on the same logit scales as the first assessment task, and thus the Rasch measures would be comparable. To identify students that would be appropriate anchors, the comparison was made multiple times using different sets of students as the anchors between the assessment tasks, and the results of the analysis of the anchored datasets were used to inform whether those anchors were acting appropriately. In all these instances it was observed that the displacement measures (the difference between the anchored measure used and what the analysis would predict if there was no anchoring) were above the significance level (≥0.50 logits). Several different methodologies were employed to identify students that were best suited for use as anchors, including: random selection, the highest achieving students, students who had closely matching ability measures in both courses before anchoring, students identified based on small displacement measures using other anchored comparisons, using the entire student cohort that is shared between both courses, using the students with abilities close to zero, and using lower ability students. The results of all these methods produced statistically significant displacement measures for all of the anchors used (values ≥ 0.50 logits), which means that there was a significant difference in the ability measure the student was anchored at and the ability measure that would have been generated for that student through the analysis. This result suggests that using the students to anchor the assessments together is not feasible, as there was no way to identify what students could be used to anchor the assessment tasks together. It is possible that this is because

the two courses are measuring different student abilities, and thus it is not possible to identify student anchors as they are performing differently between the assessment tasks.

To determine if this the case, a new approach was taken to generate comparable Rasch measures of item difficulty or student ability. The approach used was to analyse all the MCQ assessment tasks undertaken throughout the same semester in both Chemistry and Biology as a single set of data, creating a superset of the student results from all MCQ assessment tasks. The idea is that this would generate Rasch item difficulty measures for both the Chemistry and Biology items on the same scale as each other, hence providing comparable measures. The assumption when analysing the data in this way is that student ability is consistent across all the items being analysed, which means it is also assumed that the students share the same ability in both Chemistry and Biology. However, one of the benefits of undertaking the analysis in this way is that the dimensionality of the analysis can be evaluated to determine if that assumption can be justified. Applying the Rasch model to the superset can be used to generate the dimensionality using a factor analysis of the residuals present after the variance explained by the Rasch model is removed. This provides information on how much of the variance is accounted for by each contrast using an eigenvalue and the observed amount of variance that is explained by that contrast. The expected value is generated based on the amount of variance that would be expected if the contrast matches the expectations of the Rasch model, where the observed and expected should match reasonably closely – otherwise it indicates a problem with the estimation of the measures. The results of the dimensionality analysis can be seen in Table 39.

Table 39: Dimensionality Analysis Results of the Biology and Chemistry Superset

	Eigenvalue	**Observed (%)**	**Expected (%)**
Observations	163.2384	100	100
Measures	33.2384	20.4	20.4
Persons	17.668	10.8	10.9
Items	15.5704	9.5	9.6
Unexplained	130	79.6	100
1st Contrast	2.9945	1.8	2.3
2nd Contrast	2.562	1.6	2.0
3rd Contrast	2.2576	1.4	1.7

The table shows that there is more than one significant factor that is influencing the measures, as all three of the contrasts presented have an eigenvalue > 2 (the value used to signify a significant influence on student performance). The observed and expected percentages are similar for all the contrasts and thus there were no issues with the estimation of the measures; therefore the assumption of unidimensionality can no longer be made.[273,309] This not only undermines the Rasch analysis of this data, it also provides evidence that there are at least two different latent abilities that are being assessed in the Chemistry and Biology assessment tasks under consideration. If the assumption of unidimensionality cannot be justified through the analysis, it means that using the students to link the two courses cannot be justified. Even though using the superset of data does give item difficulty measures that are technically on the same scale as each other, and hence are in principle comparable, there is no reason to trust those measures to be accurate representations of the items. The inability to validate any of the assumptions that need to be made to justify this comparison mean that any results from the comparison of two independent courses are invalid, and thus this analysis cannot be undertaken unless a way to anchor the two courses together is incorporated within the construction of the assessment tasks.

While in theory it seems reasonable to expect that students have similar ability measures throughout all of their courses, and that item difficulties are comparable across multiple courses assessing at the same student level, upon reflection there is a lack of evidence to support this. This thought process is based on the expectations of the students and the courses at a university level, in that it would be expected that there are no 'easier' or 'harder' courses and the ability of students was reasonably transferable across courses if for no other reason than how it reflects the motivation and work ethic of the student, assuming that the student's motivation and work ethic is the same for every course they undertake. This is not the case in many instances, as there are many variables that can influence why students and courses behave differently. For example, it is possible that students may be enrolled within one of the courses because it is a requirement of their program, and while the other course may also be a requirement it may be of more interest to the student because of either the content or how it aligns with their desires for future studies. In this example it is likely that both the students' motivation and their work ethic will differ between the two courses, as while they may wish to do well in the course they have to undertake to complete their program, they will not be as invested as they are in the course that aligns with their personal goals. Another potential difference is the assessment strategies employed within each course, and in particular MCQ assessment tasks, as they are what is being compared. Reviewing the content and the presentation of the assessment items used within the Biology course being analysed showed that the MCQ assessment items tend to assess lower order thinking, as many of the items commonly involve recall and knowledge application. This suggests that within Biology, the MCQ assessment tasks are primarily used to assess the students' learning of the 'facts', whereas in Chemistry courses there is a significantly lower proportion of these types of items. This is because in first year Chemistry courses the MCQ assessment items are best described as 'simplified short answer response items' due to the amount of application and calculation involved in many of the items. This needs to be another consideration of the comparison, as potentially even if comparable item difficulty measures were generated, a statistically significant difference may potentially be identified simply due to different approaches to how MCQ assessments are constructed and applied within the course. What all of this suggests is even though it seems reasonable to assume that a comparison will be highly informative and be able to answer questions about the transferability of skills between courses, in reality there is the potential that this analysis will not be able to inform any of those things.

The large number of students enrolled in both courses is the reason that a comparison was thought to be possible; however, when comparing the student ability measures, it is unreasonable to anchor those ability measures only to attempt to compare those ability measures later within the analysis. Thus, for this comparison to be possible, Rasch student ability measures that can be compared need to be generated using the items to link the assessments rather than using the students as anchors. Using items rather than students removes the need to assume that student ability is the same across both courses; instead it is assumed that the item difficulty is the same across both courses, which is a more reasonable assumption to make. This does require that there be several items that are common between the assessments that can be used as anchors, and thus this analysis needs to be premeditated before the assessments are undertaken. This would require that there be some amount of overlap in course content, as it is unreasonable to include an item that is largely unrelated to the course purely for the sake of using that item to anchor the two assessments. Doing so would undermine not just the comparison but also the assessment itself, potentially giving results that are not reflective of the student ability or the item difficulty. Thus, based on this requirement it is unlikely that many courses will be able to be compared as they lack any meaningful amount of overlap in content that could be used to anchor the assessments. However, if there is no content

that is common to both courses, it is highly likely that those courses are assessing the students on a different latent trait, and thus any comparison made between those two courses would be flawed.

4.5 Conclusion

The entirety of this chapter focused around one of the research objectives, which covers the use of assessments as a tool for the comparison of student ability. Both CTT and Rasch analysis were used in all the comparisons being made; however, one key difference between the methodologies is that CTT can only be used to compare the items that are shared between assessment tasks, whereas Rasch analysis can utilise the complete assessment task as long as there is an adequate number of items shared between the assessment tasks. These comparisons relate to the research objective:

To compare item and student performance within first year Chemistry assessments over the period of a semester, across multiple years, and against Biology courses using MCQ assessments undertaken at The University of Adelaide to determine if there are any differences in performance, and if they these changes are a result of the items or the students

Each of the different aspects were addressed individually, starting with the comparison of student ability over the period of a semester, which related to the research question:

Do students show differences in their performance in MCQ assessments at different points in a semester? If so, how?

For this comparison, only two of the four Chemistry courses being analysed could be used, as they were the only courses that contained enough items shared between the lecture tests and the redeemable section within the final exam to make the comparison valid. To ensure that the student performance was not being influenced by changes in the item behaviour, the item difficulties were compared using both CTT and Rasch analysis across the different sittings of the assessments, showing no statistically significant difference using either methodology. Both methods also showed a statistically significant improvement in the mean ability of the student cohort within the redeemable section of the final exam section; however, not all the students showed improvement in their performance within the redeemable section of the final exam. Using CTT and Rasch analysis measures of student performance (raw score and Rasch student ability measure) the average change between students that improved and students that decreased in performance was roughly the same, but the larger number of students showing improvement was the cause of the overall observed improvement in performance on all 16 of the comparisons made. It was also observed that students who only undertake one of the MCQ assessment tasks tend to be of higher ability if they only undertake the lecture test, and of lower ability if they only undertake the redeemable MCQ assessment. This clearly indicates that the use of a test-retest strategy does give students the best chance to display their ability, and the final exam is not always the best place for students to demonstrate this.

The next aspect of the objective that was addressed was the comparison of student cohorts over multiple years using assessment items common to all the years being analysed, and thus relates to the research question:

Do student cohorts show differences in performance over multiple years? If so, how?

All the Chemistry courses being analysed could be used for this comparison due to the large number of items shared between years. Comparing the Rasch item difficulty showed no statistically significant changes; however, when comparing the difficulty of the items using CTT there were some deviations within the values between years that may influence the student comparison. Alternatively, the differences may be the result of the student cohort, as the values generated for the difficulty of the items are dependent upon the student cohort as they reflect the percentage of students who obtained the correct answer. The comparison of student ability using CTT showed that the 2012 Chemistry IB cohort had a statistically significant lower performance than the 2013 and 2014 cohorts. Rasch analysis showed the same significant difference, and also revealed that the 2012 Foundations of Chemistry IB cohort had a statistically significant higher ability than the other yearly cohorts it was compared to (2013 - 2015). Identifying if these cohorts are showing significant differences due to potential influences present within those years or if the cohort has a lower ability would require a deeper evaluation. The results of this research indicate that the ability of the student cohorts enrolling within first-year Chemistry courses did not decrease, nor increase, across the time period being analysed, and thus suggests that the students enrolling are similar in terms of ability every year.

The last aspect of the objective that was investigated was the comparison of student results between courses in different disciplinary areas to determine if students show similar ability levels in different courses, which addresses the research question:

Is it possible to compare student results across multiple courses from different disciplinary areas? If so, do students show similar performance across multiple courses?

Theoretically, as the Biology and Chemistry courses being compared share many students, it is possible to use those students as a reference point to place the measures within the same scale. However, the assumptions of unidimensionality and student ability being stable between the courses ultimately could not be justified by the analysis. This suggests that the latent trait being assessed is different between the courses and thus the results of the comparison would not only be unreliable, but they would also be invalid. The ability to compare between courses in different disciplinary areas would give insight into the relative difficulty of the courses and if high ability students perform consistently across multiple courses. Theoretically the comparison is still possible; however, thought would need to be given to whether the two courses are worth comparing based on the information that would be learnt and how that can be applied. It is also important that if items are used to link the assessment tasks that the shared items can be included within both assessment tasks without invalidating the assessment task through unrelated assessing of the students.

Chapter 5: Methodology of Assessment Analysis

5.1 Section Outline

5.1.1 Research Questions

The most important aspect of assessment analysis is that any assessment task used to measure student competency is analysed to ensure it is fulfilling its purpose. Therefore, the fact that an analysis is undertaken is more important than the actual methodology used to conduct that analysis, as any methodology will provide more information than no analysis will. However, if a decision needs to be made about which methodology will provide the most relevant information to the purposes of the analysis, it is important to consider the options available, which is the focus of one of this project's research questions.

What is the most appropriate way to analyse MCQs in order to provide an approachable methodology that can be used to improve assessments?

Throughout this research two methods of MCQ assessment analysis have been utilised, Classical Test Theory (CTT) and Rasch analysis, and thus the application of those two methods can be explored within this research. Comparing the information that was obtained, and accounting for the knowledge and time that is required to obtain that information, can inform which methodology is most appropriate for improving MCQ assessment tasks.

5.1.2 Project Objectives

To answer the question of what is the most appropriate methodology, the main considerations need to be how approachable that methodology is and how effective that methodology is at obtaining the desired information. Thus, the project objective related to answering this question can be stated as:

To identify the most approachable and effective methods to analyse MCQs, and develop a process that can be used to improve any MCQ assessment

The most approachable methodology and the most effective methodology are not necessarily the same, and thus this refers to the most approachable and effective method to fulfil the purposes of the analysis. Thus, there is never going to be a 'one size fits all' approach to assessment analysis, and instead an informed decision needs to be made based on what information is desired and the methods that can be used to obtain that information. Therefore, this objective is more about informing others about the possible application of different methodologies rather than giving them a structured method that they can apply themselves. That is not to say that a structured methodology is not possible to produce, but rather there is no way that it can fulfil the needs of every analysis and thus it is more important that assessors are able to make their own educated decision about the most appropriate methodology for what they require.

5.2 Comparing the Results of Assessment and Item Analysis

5.2.1 Differences in Assumptions

To apply either Classical Test Theory (CTT) or Rasch analysis, the assumptions of these methodologies need to be met by the data being used, which may influence which method is the most appropriate for analysis. The assumptions of CTT are considered 'weak assumptions', which means that they are easily met by most datasets, and can be summarised as:[266]

- A student's observed score is equal to their true score plus their random error
- A student has an expected random error of 0
- Across the student cohort the average random error is 0
- There is no correlation between a student's true score and their random error
- Random errors across multiple tests are uncorrelated

All these assumptions address the theory that the results of an assessment reflect the true score of the student and some amount of random error that cannot be measured but averages to 0 over multiple assessments. The random error may be due to factors such as student guessing. mistakes, or anything that may have influenced the students to deviate from their true score. The random error is expected to shift the observed score up and down - for example in some assessments guessing may work in the students favour and in others it will not - but over time with enough assessments it is expected that the random error will average to 0 and the student's final grade will reflect their true score. It is also expected that across the student cohort the random error will average to 0, as this means that the assessment and the items do not include a random error factor when assessing their performance. An important note is that all of CTT's assumptions are focused on explaining the results of the students, which means that the results of CTT are more reflective of the student cohort than the assessment task and items.

In comparison to this, Rasch analysis is built around stronger assumptions that have requirements the dataset must meet before the Rasch model can be applied, which are:

- Each person is characterised by an ability
- Each item is characterised by a difficulty
- Ability and difficulty can be expressed as numbers on a line
- The difference between these numbers, and nothing else, can be used to predict the probability of observing any scored response

Not mentioned explicitly within these assumptions is that as each assessment is characterised by only two measures (student ability and item difficulty) making these measures unidimensional, but this also means that there is no factor included for student guessing within assessment tasks nor is there a measure of item discrimination. While these stronger assumptions are harder to justify than the assumptions that CTT makes, the reason that Rasch analysis requires these assumptions to be met is because they are requirements for producing independent measures. If these assumptions cannot be met by the data it means that the data is not suitable for producing independent measures of student ability or item difficulty, and thus there is no reason to apply the Rasch model to that dataset as it will provide no meaningful information.

The difference in the strength of the assumptions is one of the reasons that CTT is seen as the more approachable method for data analysis; however, the strength of the assumptions also influences the outcomes that can be expected from each methodology. One of the major benefits of Rasch analysis is that it produces independent measures for both student ability and item difficulty that are expressed on the same relative scale, which can only be done due to its assumptions. It is important

to consider the assumptions of each methodology before they are applied because not every dataset will be applicable for Rasch analysis; however, most datasets can be reasonably expected to be applicable for CTT analysis. That being said, just because it is believed that a dataset will not meet the assumptions of an analytical methodology does not mean that the analysis should not be undertaken, as in many cases it cannot be known whether the assumptions are violated until after the analysis is completed. For example, the hardest assumption to satisfy within Rasch analysis is the assumption of unidimensionality; however, as what each dimension represents is not explicitly known, it cannot be judged whether unidimensionality is met until the analysis has been performed. Therefore analytical methods should not be ignored because their assumptions are harder to meet, but rather thought and consideration should be placed into how those assumptions influence the outcome of the analysis and what is required to fulfil the purpose of the analysis.

5.2.2 Information Obtained by Each Approach

The information obtained by both methodologies about the assessment task and items share a large amount of overlap in what they inform; however, there is a difference in the nuance that the information provides. CTT can provide a range of different measures that are related to the assessment task; in this research KR-20 and Ferguson's Delta were chosen. KR-20 is a measure of internal consistency that indicates whether the performance of the student cohort is consistent across the entirety of the assessment task; using this measure allows evaluation of whether the assessment task is assessing the students on related concepts and ideas. Ferguson's Delta is a measure of the discriminatory power of the assessment task; using this measure allows evaluation of whether the assessment task can differentiate between students of varying abilities. In comparison to this, Rasch analysis gives three different measures that can be used to judge an assessment task. The first measure that needs to be evaluated is the assessment's dimensionality, because if the assessment contains multiple dimensions that influence student outcomes it means that the assumptions of the Rasch model are not maintained, and therefore Rasch analysis cannot be used. If the dimensionality results are within expectations, it means that the assessment task is consistent in what student trait is being measured. The other two measures generated by Rasch analysis are the reliability and separation of that assessment task, which relate to the reproducibility of the measures and how well the measures can be differentiated from each other respectively. The reliability and separation measures are given for both the student cohort and the items utilised, which means that the performance of the assessment task can be evaluated based on its ability to measure student ability and how well the items contribute to fulfilling the purpose of the assessment task.

Evaluating an assessment task using CTT or Rasch analysis will produce semi-analogous measures, as both give measures of content uniformity (KR-20 and dimensionality) and how well the assessment differentiates between varying abilities (Ferguson's Delta and separation). However, in addition to those measures Rasch analysis also provides information on the expected reproducibility of the assessment results and thus the expected consistency of the student and item measures (reliability). Neither method of analysis provides any diagnostic information to resolve any issues identified within the assessment task, and thus both methodologies require deeper analysis of the individual aspects of the assessment to determine what may be causing any issues found. Rasch analysis does provide slightly more information about the potential origin of the issue due to it providing information on both the items and the students within the assessment tasks, whereas CTT does not separate these factors. The reason the Rasch can separate these factors is due to its ability to calculate independent measures of student ability and item difficulty whereas CTT cannot give

independent measures and thus is unable to separate the two factors due to their dependence on one another.

When the individual items are being analysed, CTT readily provides three different values to describe each item: item difficulty (percentage of the student cohort obtaining the correct answer), item discrimination (comparison of top and bottom quartiles of students to ensure the item is separating students of varying ability), and item correlation (how well the results of this item correlate to the overall results of the assessment task). To determine which items are appropriate based on these measures means that a series of threshold parameters are referred to, or alternatively the measures may be justified based on the expectations and the purpose of the assessment task. In contrast, Rasch analysis provides the item difficulty measure and a series of measures that inform how well the item fits the expectations of the model due to its confirmatory nature. These measures are: the item infit (measures significant variation from students whose ability is close to the item difficulty), item outfit (significant variation in item answers across all students), and observed vs. expected (comparison of the observed results to the expected outcome based on the Rasch model). As these measures are used to compare to the expectations of the model the extent of the items deviation is used to define when an item is classified as 'problematic', which can be confirmed not to be an outlier using significance testing.

When evaluating the performance of individual items, the two different methodologies vary in their expectations of the items, and thus this has a large influence on how an item is interpreted by the analysis. While both methods give an item difficulty, what that measure represents is very different between the two methodologies. Within CTT the item difficulty reflects how many students correctly answer the item, which is used to represent the difficulty the item posed to the students. The issue with this approach is that it is heavily influenced by the student cohort, and thus the item difficulty is just as reflective of the student cohort as it is of the item itself. Within Rasch analysis, the item difficulty measure is relative to the other items within the assessment task, as the mean item difficulty is made to be the 0 on the logit scale, and thus without the context of the rest of the assessment the item difficulty only provides vague information. However, when analysed alongside the other items present the item difficulty clearly shows the relative difficulty of each item independent of the student cohort, and the difference in the difficulty measures of two items is expected to be constant in all assessment tasks they are both included within. There are two major differences in the item difficulty measures produced by each methodology: the first is the influence of the student cohort on the results, and the second is the transferability of the information across assessments. Despite these differences both measures are used to evaluate whether the item matches the purpose of the assessment and if the student cohort performs as expected based on these measures. Within CTT this evaluation uses item difficulty threshold parameters of 0.30 – 0.70, whereas in Rasch this evaluation is performed by comparing the item difficulty to the average student ability to determine if that item is providing any meaningful information about the student cohort. Thus, by itself, the item difficulty is not a good diagnostic tool for determining if the performance of an item is reasonable within an assessment.

Item discrimination and item correlation measures in CTT are used to supplement the item difficulty measure to evaluate item performance, whereas Rasch analysis compares the item difficulty generated to the expectations of the model. Item discrimination is used to evaluate if the item can differentiate between high and low ability students, which ensures that the item is not subjected to large amounts of guessing or misinterpretation. In comparison to this, Rasch does not have any measure that addresses item discrimination or the influence of guessing; instead, it assumes that

these values are equal throughout the entire assessment. It is worth noting that it is possible to use Rasch analysis to generate an item discrimination; however, it is done as a post-hoc analysis. The issue with CTT's discrimination measure is that it can fluctuate based on item difficulty, and thus the item discrimination needs to be considered alongside other item measures provide context. The item correlation ensures that what the item is assessing matches the rest of the assessment; however, assuming the assessment was reviewed before its use, it is unlikely that items will not correlate unless it is expected that they are assessing unique content. Rasch uses item fit measures of infit, outfit, and observed vs. expected to determine if the item is significantly deviating from the model's expectations (measured through the use of significance testing), where the extent of its deviation, and the direction of any deviations can be used to inform how an item is performing. Thus, while both methods are analysing the items on similar constructs, the breadth and nuance of the information given by Rasch analysis provides more information on how an item may be deviating from expectations. This is particularly true when graphical representations are used to show the performance of the item in comparison to the model's expectations, something that cannot be done using CTT, as this can highlight where any issues are occurring and if they need to be addressed or simply represent random variation. The breadth and nuance of this information may be superfluous for some analyses, particularly if the assessment has been previously evaluated, as CTT is capable of highlighting any items with major issues and further analysis can be done to supplement that result afterwards if deemed necessary. The different measures generated by each methodology, and the different areas that they can be applied within are summarised within Table 40.

Table 40: The Comparison of Classical Test Theory and Rasch Analysis in how they Approach Different Aspects of Assessment Evaluation

	Classical Test Theory	**Rasch Analysis**
Assessment Correlation How well the results of the assessment reflect the students' ability within the topic being assessed	KR-20	Dimensionality
Assessment Separation The ability for the assessment to differentiate between the results of the students and the items	Ferguson's Delta	Student/Item Separation
Assessment Reliability How reproducible the results of the assessment are	Not possible	Student/Item Reliability
Item Difficulty How the analysis classifies the results of the items	% of Students Answering Correctly	Independent Logit Measure
Student Ability How the analysis classifies the results of the students	Raw Score	Independent Logit Measure
Item Discrimination How well an item differentiates between students of high and low ability	Top Quartile Compared to Bottom Quartile	Assumed Constant across all Items
Item Correlation How well the results of an item reflects the ability of the students within the entire assessment	Compared to Overall Assessment Result	Compared to Model's Expectations

Guessing Factor The inclusion of a way to account for students guessing the correct answer within a MCQ format	Can be Applied Separately	Assumes Net Benefit of 0
Deviations from Expectations How the model detects items/students that are not functioning correctly, and where that deviation is occurring	Threshold Values, Rationalise Expectations	Infit/Outfit MNSQ and ZSTD
Distractor Analysis How the model breaks down the options selected within a MCQ assessment	Post-hoc analysis	ICC with options selected relating to the ability of the students
Student Evaluation The ability for the model to analyse the results of the students	None generated	Same as Items
Differential Item Functioning Comparing if different student groups react differently to specific items	Post-hoc analysis	Any Factor Included within Analysis to Model's Expectations
Differential Person Functioning Comparing if different items groups react differently to specific students	Post-hoc analysis	Any Factor Included within Analysis to Model's Expectations
Assessment Linking Being able to link two or more assessments such that their results can be compared	Only Shared Items	Entire Assessment Assuming ≥ 5 Items can be Used as Anchors
Item Comparisons Being able to compare the results and analysis of items to each other	Influenced by Students	Compare on the Same Logit Scale
Student Comparisons Being able to compare the results and analysis of students to each other	Influenced by Items	Compare on the Same Logit Scale

The major advantage of the Rasch model for item analysis is its ability to generate an independent measure of item difficulty, as the dependence of CTT's measures on the student cohort means that the expectations of an item also need to be considered within its analysis. For example, a relatively easy item might show problematic item difficulty and discrimination measures but reasonable correlation using CTT that might raise concerns for the item; however, if the item is assessing a basic but important concept, then even though its measures may lie outside the accepted thresholds the item may still be performing as expected. If the same item is analysed using the Rasch model, it will be acknowledged that the item difficulty is relatively low and expectations of the Rasch model are adjusted to compensate for that, and thus it is reasonable that Rasch analysis would have no issue with this item. Thus, even though the confirmatory nature of the Rasch model might be initially seen as an issue because each item performs differently, Rasch analysis can account for those differences based on the item's difficulty measures. In comparison to this, the fact that CTT is a student focused method of analysis means that the item measures are just as reflective of the student cohort as they are of the items themselves. Therefore, within CTT there always needs to be a check of whether the

item measures are expected, whereas in Rasch the expectations of an item are shifted based on the item itself.

5.2.3 Comparison of the Analysis Results

The results of the assessment analysis from both methodologies used in this research showed that the assessments had almost no issues except on occasion the KR-20 from CTT analysis was slightly below the expected threshold, and the student separation and reliability from Rasch analysis was also low in some assessments. However, these issues can be identified as being caused by the small number of items included within the assessment tasks, which results in less information being gathered about the students. Thus, both methodologies suggest that there are no obvious underlying issues within the assessment tasks being analysed in this research based on the measures they generated for the assessment tasks. The only real difference between the two methods of analysis is the amount of information that they provide as to why there are no obvious issues within the assessments.

By comparing Appendix 7.8 and Appendix 7.9 it is clear that Rasch analysis identified far more problematic items throughout the assessments being analysed than CTT did (83 items compared to 12 items, where 41 problematic Rasch items were determined to represent major problems compared to 4 major problematic items identified by CTT), and it is important to consider why that is the case, as in theory both of these methods should be identifying the same problematic items. It should also be noted that even though Rasch analysis identified significantly more items, there were 2 items that were identified to be problematic by CTT that were not seen as problematic by Rasch analysis; however, both of these items represented minor problematic items and are likely only identified based on the difference in the expectations of the methodologies. One reason for the large difference in the items identified is that CTT will never identify overfitting items as an issue, which are items that highly discriminate between students and at their most extreme no students below a certain ability will correctly answer the item and above that ability all of the students will correctly answer the item. The reason that these items are not identified by CTT is that there is no upper bound on the discrimination value, and higher discriminations are thought to represent higher quality items, as they are able to clearly differentiate between students. The reason that overfitting items are treated as problematic within Rasch analysis is because they are too predictable and give no new information about the ability of the students. If an assessment task is constructed of only overfitting items, the CTT measures would indicate that the items are performing exceptionally well; however, all that would be learnt from that task is the bottom half of students and the top half of students, and there would be no way of differentiating between the students within those groups. Therefore, the problem with overfitting items is the lack of information that they provide the assessors, which is less of a problem than underfitting items which are prone to guessing and misinterpretation, but are still capable of causing issues for generating student measures. This difference accounts for 40 of the items identified by Rasch analysis that were not identified by CTT (25 minor, 1 potentially major, and 14 major), which leaves 33 problematic items that need to be explained.

Another reason for the difference in the problematic items identified is due to the difference in the requirements for classifying an item as problematic. Within this research the Rasch analysis thresholds used were recommended for high-stakes assessments, whereas there were no adjustments made to the CTT thresholds for adequate MCQ performance. The harsher guidelines used in Rasch analysis likely resulted in several items being identified as problematic that otherwise would not have been (e.g. minor flawed items). The reason that the harsher thresholds were used

was due to the inclusion of the number of times an item appears problematic as a factor that could be considered. The use of multiple assessments acts as a filter to ensure that items were not simply being identified as problematic due to random variation within one year, but rather that they lie outside of those thresholds consistently. In comparison to this, the thresholds for CTT remained unchanged due to their dependence on both the student cohort and the item itself. What this means is that these measures are far more prone to variations, and often the threshold values need to be considered on an item by item basis as some items can be justified to lie outside of the regular thresholds depending on the expectations of the item. Thus, there is no reasonable adjustment that can be made to the CTT threshold values that would hold true for all the items, as each item that surpasses the threshold values needs to be individually evaluated to verify if the item is performing outside expectations. This means that problematic items identified through the use of CTT require an additional level of analysis before the item can be considered to be problematic, whereas the use of the thresholds within Rasch analysis is enough to justify labelling an item as problematic.

The third reason for the difference stems from the amount of information each methodology provides to justify classifying an item as problematic. For example, using CTT it is reasonable that an 'easy' item (high item difficulty) has a poor discrimination, as if all of the students are answering it correctly then the item cannot differentiate between the high ability and low ability students. The correlation is likely to be reasonable for an item of this nature, which means that potentially one or two of the item measures lie outside their recommended thresholds, but despite this there is the potential that the item is not classified as problematic. It could be considered problematic, reviewed, and determined that it is expected to be an easy item and thus the results observed are expected. Another factor that skews these results within CTT is that the discrimination measures are often only approximations due to issues in separating students by quartiles within assessment tasks that only have a limited number of outcomes. What this lack of information means is that there needs to be more leeway in what identifies as a problematic item within CTT than there is within Rasch, which means that items that may be problematic are treated as though they are constructed in that manner deliberately and thus are left unchanged.

The large difference in the number of problematic items identified may initially be cause for concern, as ideally both methodologies should identify all the same items, but the differences in the methodologies can account for these inconsistencies. This highlights the importance of choosing the most appropriate methodology in each circumstance, as CTT can identify problematic items that are the largest threat to assessment validity but will not give the nuance of a method such as Rasch analysis. In some instances that nuance is important, particularly in high stakes assessments, but in other cases it may be less relevant. Thus, each assessor needs to determine the expectations and the standards that their assessment needs to conform to as this will help them to decide the most appropriate methodology to fulfil their requirements.

In comparison to what is observed for each methodology when comparing problematic items, the opposite occurs when identifying gender biased items, as observed within Appendix 7.12 and Appendix 7.15. There were 27 items identified by CTT and 14 items by Rasch analysis that showed consistent differences in the performance of male and female student cohorts, where only 8 items were identified by both methods. It is important to remember that the method used to identify gender biased items using CTT utilises that assumptions of CTT but is a chi-square test comparing the results of male and female students within the cohort, where it is assumed that both genders are equally as likely to provide the correct response. The fact that there were more problematic items identified by Rasch analysis may explain why there are less gender biased items observed, as any

item identified to be problematic was excluded from this analysis. This is because if an item has underlying issues then it is unreasonable to expect that item to perform in any logical way, and thus gender bias may be introduced due to its problematic nature. While the items identified by CTT as problematic were also removed for the same reasons, the fact that there were fewer items identified as problematic meant that fewer items were excluded from the analysis. The largest difference in the comparison is that CTT must assume that male and female students are equally as likely to give the correct answer; however, Rasch analysis does not have to make that assumption due to the independent ability measures it generates. This means that Rasch analysis can account for differences within the student cohorts due to factors such as self-selection whereas CTT cannot.

The last item analysis conducted within this research was ensuring that they were stable across multiple assessments to allow for the comparison between student cohorts that shared items across their assessments. To make this comparison the item difficulties produced by both methodologies had to be compared across different time intervals (see Section 4.3.2 and Section 4.3.3) using a paired sample t-test to ensure that the difficulties do not significantly change between those time intervals. Even before the comparison is made it can be concluded that Rasch analysis is a much better method for comparing between assessments, assuming there is some way to link the two assessments together, as the independent item difficulties produced make the items comparable between different student cohorts, whereas the CTT item difficulties are dependent upon the student cohort and thus different student cohorts may influence those results. It was observed that in all the comparisons made where the assessments could be linked through overlapping items that Rasch item difficulty measures showed no significant differences. In comparison to this, as soon as the student cohort showed a large amount of variation (i.e. when student cohorts from different years were compared), issues started to arise in the CTT item difficulty comparison. This then leads to concern around the validity of the comparison being made as potentially comparing the results of the student cohort will not be representative of differences between the two but due to differences in their assessment experience. Using CTT minimises the number of items that can be compared, as only items that are present in all the assessment tasks compared can be used as there is no way for CTT to account for item differences. Conversely, so long as there are enough items that can be used to link the assessments (i.e. usually a minimum of 5 items) the entirety of an assessment can be compared using Rasch analysis regardless of how many items are different between the two assessments due to the independence of the student ability and item difficult measures. What this means is that even though for all the other item comparisons both CTT and Rasch produce semi-equivalent measures for the purposes of comparing between assessments, Rasch is the superior method due to the independence of its measures.

5.2.4 Addressing Issues within Assessment Tasks and Items

While both methods can be used to identify problematic items within assessments, another consideration needs to be what the methodologies offer once the item is identified. The use of CTT in this research gives 5 values that could be used to identify how the assessment task and the individual items were performing and what may be causing the issue (see Table 40). In comparison to this, Rasch gives 4 separate measures for the performance of the assessment task, dimensionality information, and 10 different values for each individual item that can be used to evaluate their performance (see Table 40). In addition to this, Rasch analysis has the ability to generate various graphs (e.g. Wright Map, ICC, DIF plot, etc.) that can be used to obtain different perspectives on the assessment task and item fit. There are other ways of approaching the data obtained through CTT; however, these options are limited due to the amount of information that CTT provides.

The wealth of information generated through Rasch is vast; however, it is important to utilise that information in a way that improves outcomes of the assessment task. If the objective of the analysis is to ensure the functionality of the assessment and identify any problematic items that the assessment task contains, this is where the two methods of analysis are the most comparable in terms of utilising the information they provide. The measures that allow for the evaluation of the assessment tasks from both methods ensure that there are no large underlying concerns within the assessment task itself, and usually any concerns identified relate to either the number of items or the number of students associated with the task. If Rasch analysis is being used to evaluate the assessment task, an issue with the number of items will appear within the student measures and if there is an issue with the number of students it will appear within the item measures. Even though CTT does not separate students and items in its assessment task measures it is usually reasonably easy to determine if there is a lack of students and/or items by evaluating the number associated with each. When evaluating the assessment items all three of CTT's item measures need to be evaluated (difficulty, discrimination, and correlation); however, within Rasch there are only two key values that need to be evaluated, the infit and outfit values. It should be expected that the most problematic items within the assessment are identified by both analyses (disregarding overfit items which are not seen to be an issue by CTT), but within CTT after the problematic items have been identified there is no information generated on how they can be improved, whereas in Rasch there are other values generated for this purpose. For example, if an item is showing a high item difficulty (i.e. a large percentage of students obtain the correct answer) and a low discrimination and correlation value it suggests that for some reason higher ability students are not selecting the correct answer as often as expected when compared to lower ability students. However, there is also the potential that the ease of the item is skewing the results in such a way that the item appears worse than it is. In comparison to this Rasch analysis gives information on where the issue is occurring through its infit and outfit measures, as well as the significance of that deviation from the expected values which means that there is a high degree of confidence in the conclusions generated through Rasch analysis.

The next step is to break down the item construction and determine how any problems within the item can be fixed. As previously discussed, (Section 3.3.4) any problems within an item may occur within either the stem or the options, and there are no analytical techniques that can quantify issues within the stem. Therefore, the only way these methodologies can contribute to the item breakdown is through distractor analysis, which can be done to differing extents using CTT and Rasch analysis. The extent of a CTT distractor breakdown is quantifying how many times each option is selected within the item and the average raw score of the students selecting that option; however, not only is Rasch also able to do this, it also identifies the ability level of the students that are selecting each option and if that matches the expectations of the Rasch model. This means that Rasch analysis provides more confidence in identifying any issues within the options, as the breakdown of how student performance matches option selection can be used to pinpoint areas of concern.

When using these methodologies to analyse assessment tasks and individual items, in general CTT broadly identifies areas that are of the largest concern, whereas Rasch analysis will be able to identify those areas in addition to more nuanced issues within the assessment. This means that while both methods have the potential to improve the assessment, CTT will only address the most obvious problems within an assessment. However, sometimes that is all that is required from assessment analysis as not every assessment task needs to be heavily scrutinised. For example, a high stakes assessment task that influences a student's academic future should be analysed

thoroughly to ensure that there are no issues within the task that will unfairly influence the students' results. In comparison, any low risk or practice assessment tasks simply need to ensure that the assessment and its items are not behaving erratically and allow the students to demonstrate their ability without any unrelated influences. Thus, for a high stakes assessment, Rasch analysis or another detailed methodology should be employed to evaluate the assessment, but for a low risk assessment, CTT is a perfectly reasonable method to use.

5.3 Methodology for Analysing Student Performance

5.3.1 Assumptions and Accounting for Them

Often systematic analysis focuses on the assessment task and the individual items, but there is also the potential to analyse the performance of the students. As the analytical methods being used are still CTT and Rasch analysis, the assumptions that were made when analysing the assessments and the items are expected to be true when analysing the students. As student ability is the measure of interest, this means that there is less concern about whether the students are adversely influencing the outcomes of the assessment. That does not mean that there cannot be outliers within the student cohort, as there likely will be students who perform outside of expectations, but those outliers will not influence the results of other students. CTT assumes that a student's result is the combination of their true score and random error; however, CTT does not analyse the students individually and only considers the results of the entire student cohort. Rasch modelling can analyse the entire student cohort and the individual students, where individual students deviating from the model suggests that the student is performing unexpectedly and therefore may either be guessing or utilising 'test wiseness' within the assessment. The amount of variation seen within the student cohort is influenced by the number of students, the number of items, and the difficulty of those items relative to the student ability, and thus these are the first factors that should be considered when analysing any variations observed. It is reasonable that students show some amount of misfit because of random variation, and thus individual students deviating from the model should not be treated as problematic in most instances.

5.3.2 The Quality of the Information Obtained

If students deviating from the expectations of the analysis are not treated as an issue it does suggest that there is little interest in the performance of the students from an analytical point of view; however, this stance is dependent upon the purpose of the analysis. If the purpose of the analysis is to evaluate the assessment task and the items to ensure that their results are reflective of student ability, then there is no reason to analyse the students themselves. If the purpose of the analysis is to identify individual students that may be struggling with the content, break down the distribution of student abilities, or potentially identify students that may be gaming the assessment then analysing the students is important to fulfil those objectives. In instances where analysing the individuals is important, CTT should immediately be discounted as an analytical approach, as it does not consider the results of the individual (see Table 40). This means that the only method used in this research that could provide information on the individual students is Rasch analysis. The student analysis provided by Rasch analysis is robust in that it can be used to fulfill any of the purposes described earlier; however, depending upon the number of individuals being analysed, this can be a time-consuming process to consider the results of each individual student. While it is possible to undertake this process, it should only be done if there is a predefined purpose for the student analysis rather than using it to look for issues within the student cohort. This is because if the student cohort is large enough, eventually there will be at least a handful of students who perform

unexpectedly due to the random variance present within an MCQ assessment, which means that the expectations of the analysis need to be shifted to account for the amount of variation that can be expected. The most important aspect whenever the students are analysed is that the purpose of the analysis is clearly stated, as otherwise there is no point in breaking down the student performance.

If the goal of the analysis is not to analyse the results of individuals, but rather to make a comparison between individuals or cohorts, then there are other potential approaches. It is important that there is a defined purpose for the comparison and that the assessors are aware of potentially influential factors so they can be accounted for throughout the analysis. For example, when the gender bias analysis was undertaken using CTT it was done as a comparison of the male and female student results on each item assuming that both male and female students have an equal probability of obtaining the correct answer. While this assumption appears reasonable there are potential reasons why that may not be true (self-selection for example, as discussed within Section 3.4.2) and that possibility needs to be considered when evaluating the results. Conversely, Rasch analysis can account for any differences in student ability between the two cohorts due to the independence of the student ability measures. A different concern that needs to be addressed whenever the results of two different assessments are compared is whether the changes observed are solely due to changes between the student cohorts and not the result of the any outside influences. Using Rasch analysis this issue is easily addressed, assuming both assessments can be placed on the same scale, due to the independence of the measures. Using CTT, the assumption that only the student ability can influence the outcome of the assessment (assuming the same items are used) is much harder to justify due to the dependence of the item difficulty on the student results, which means that the assumption may need to be validated logically rather than statistically, which is not ideal. Thus, the quality of these analytical methods is dependent upon their ability to account for the assumptions and influences that affect any evaluation or comparison, which is better performed by Rasch analysis than CTT due to its independent measures.

5.3.3 Comparison of Student Performance Analysis Results

Within this research, student analysis was used to compare male and female student ability (Section 3.4), the results of students in two overlapping assessment tasks (Section 4.2), and yearly student cohorts enrolled in the same course (Section 4.3). When male and female students were compared it was to ensure that the analysis of items for gender bias was not being influenced by differences in the male and female student cohorts. CTT used the students' raw scores, whereas Rasch analysis used the student ability measures, and despite this difference there is a reasonable amount of overlap between the cohorts identified to have significant differences in male and female student ability. Both methods identified that 15 of the student cohorts that undertook each of the assessment tasks being analysed (out of a total of 64 assessment tasks that were considered) had significant differences in male and female student ability, where 9 of the student cohorts were identified by both methodologies. Therefore, both methodologies identified 6 student cohorts that showed significant differences in male and female student ability where no difference was observed within the other method. This does indicate that despite the differences in the methodology there are similarities in the results identified; however, this should not be entirely unexpected. Even though Rasch student ability is not dependent upon the student's raw score, there is a large correlation between a student's raw score and their relative placement within the logit scale, and thus it should not be unexpected that there are similarities in raw score comparisons and ability measure comparisons. Even though the results are similar, the independence of the Rasch student ability measures means that more confidence can be placed in the results of Rasch analysis. For example, the purpose of this comparison is to ensure that there are not pre-existing differences in

male and female student ability and thus the assumption that both will perform equally well is required for a CTT comparison of the gender results on individual items. However, if there were gender biased items within the assessment then it stands to reason that this will influence the raw scores of the students, and thus the significant difference in male and female student cohorts may be reflective of the items rather than the students. Then, when the analysis of the item takes place, there is the potential that gender biased items identified will be ignored due to the significant difference seen within the student cohort, but there is no way of separating whether the bias is the result of the students or the items. In Rasch analysis, the independence of the measures means this is not a concern, as it can be clearly determined what is influencing the outcomes. This is a trend that is echoed throughout all of the student ability comparisons, and while CTT does produce reasonable results there is always the potential that different influences may be distorting the results.

The comparison of student results in a test-retest assessment where the students took an almost identical assessment later in the year showed very similar results using both CTT and Rasch analysis. All the student cohorts analysed were identified to be significantly higher performing within the second assessment (redeemable exam section), and both methodologies also had the nuance required to demonstrate that even though the change was significant not all the students were improving. This indicates that in this example CTT functions in the same way as Rasch analysis and can give the same results without having to undertake the deeper level analysis. The issue is that compromises had to be made within the CTT comparison that were not required of Rasch analysis, most notably that only the overlapping items could be compared in CTT. While in this instance this compromise did not influence the outcome of the analysis, it does become a greater concern for assessments that contain fewer overlapping items. This is because when too many items are removed from the assessment results it impacts the validity of the comparison as it disregards aspects of student competency being measured in the assessment. Due to the independence of the student ability and item difficulty measures within Rasch analysis, no items need to be excluded from the comparison as long as there are enough overlapping items to link the assessments together.

Comparing across yearly student cohorts also showed a great deal of overlap in the significant differences identified within CTT and Rasch analysis, but also highlighted some of the issues with CTT's assumptions. Within this comparison the item difficulty was found to vary significantly between years using CTT, whereas there was no such issue identified by Rasch analysis. Thus to undertake the analysis it needs to be assumed that the significant changes in item difficulty are a result of the student cohort rather than the items themselves, as otherwise any significant difference in the student scores may not be reflective of changes in the student cohort but rather due to changes in the items. While the assumption that it is the students and not the items that are changing is reasonable, there is no way of conclusively showing that this is not a concern within the analysis. In addition to that, while the years that showed significant differences in item difficulty did have changes in the ability of the student cohort, most of those differences were not found to be statistically significant. Then there is the question of whether the changes in the student cohorts were large enough that they may have influenced a significant difference in the performance of the items, which cannot be answered by any analysis. Another issue identified that is semi-unique to these assessments is that when optional assessments are present, the raw score of the students over all the assessments presented cannot be used as the students that undertook more assessment items would be expected to have a higher overall score. Thus, the student percentage in MCQ assessment items was used, and while this does solve the issue of the students having different

potential maximum scores, it does cause concerns about what assessments the students undertook and whether this may influence the student's results. It has to be assumed that across the entire student cohort there are enough students taking the same or similar approaches to the assessment that any influences balance out. None of these issues are a concern within Rasch analysis due to the independence of the measures and its ability to give measures that can be placed on the same scale regardless of whether or not the students have completed the same number of items. Despite these concerns and concessions that had to be made when utilising CTT, the results produced by both methods show large amounts of similarity for the same reasons as discussed when making the test-retest comparison. However, the large amount of concessions that are required to utilise CTT means that more confidence can be placed in the results of Rasch analysis as these concessions were not required to make the comparison.

In many cases there is no reason to analyse the performance of the students within assessments, but the importance of producing independent measures within any comparison cannot be understated, as without independent measures there will always be some doubt as to which factor is influencing the outcomes observed. The ability to produce measures of student ability and item difficulty that are relative to each other, regardless of the differences in the circumstances, is also extremely helpful in any student comparison being made. However, it is the lack of any way to analyse individual students and the assumptions required by CTT that should make Rasch analysis the preferred method for analysing student performance.

5.4 What Methodology, and When?

5.4.1 The Application of a Methodology

Before any methodology is applied it is important that the purpose of the evaluation is well defined, as this will influence what the most appropriate methodology is and will help clarify the important aspects of the evaluation. The factors that need to be considered when deciding the purpose of the evaluation are deliberations such as the investigation of individual students, whether overfitting items are a concern, the approach toward problematic items, potential for comparisons between assessments, and the stakes of the assessment task for the students. These factors need to be considered because they will change the information that is required from the evaluation. In general, if large amounts of information are desired about the assessment and the individuals, a more detailed methodology such as Rasch analysis is required. If the evaluation is being used to identify major concerns within the assessment and nothing more, then a method such as CTT might be the best approach. There may also be factors that need to be considered that lie outside of the assessment analysis such as any time restrictions on the evaluation, or concerns about the knowledge required to apply different methodologies. While these are legitimate factors that may influence which methodology is used, it is important that they are not the sole influence for why a method was chosen, and if they do have an influence on the method chosen it is important that the purpose of the analysis is also re-evaluated to ensure that the methodology chosen is capable of providing the desired outcomes.

One major factor that is a detractor for more detailed methodologies is that their additional complications are considered to make them a more time-consuming way to evaluate assessment tasks. While the theory behind these methods may seem more complicated, once it is well understood the time aspect of the evaluation is only relevant for the most basic investigations, and even in that instance it can be just as quick to perform a Rasch analysis as it is to apply CTT. If a more

detailed evaluation is being undertaken with the intent to identify all problematic items and the issues causing them then it may become more time efficient to use a methodology such as Rasch analysis rather than CTT. This is because the extra information provided by Rasch analysis saves time due to the confidence it can give assessors in the results and the additional information it supplies regarding the underlying causes of any issues. Thus, while the initial time investment into the methodology may be higher, the amount of time saved means time efficiency should not be the driving force for choosing any one methodology.

Before any analysis can begin the data needs to be prepared in a format suitable for the method of analysis chosen. CTT simply requires the results of each student on each item (i.e. whether they answered each item correctly), which ideally is presented in a tabulated format as this makes the data easy to manipulate. Rasch requires the same information, although ideally the options that the students selected are also included, and it also requires information about the assessment itself to be included within the datafile to be analysed. This additional information includes the number of items and students within the assessment, the answer key, acceptable student responses, student and item labels, and any specific functions the analysis is expected to undertake – most of which is also required to be known when applying CTT. There are numerous other factors that can be included within the information such as specific student groups (e.g. gender, age, etc.), how to treat incorrect responses or invalid responses, and how the information should be presented as a few examples. This does mean that more information needs to be included within Rasch analysis and it is required to be specifically formatted for the program undertaking the analysis. In this research Winsteps[308] was used to conduct the Rasch analysis, which requires deliberate construction and formatting of all of the relevant information before it can be uploaded to the program. In comparison to this, all the CTT analysis was undertaken within Excel so only minimal additional formatting was required before analysis could begin, as this was the format in which the raw data was received.

The methods that are available to be utilised to analyse the results of an assessment should also be considered, as Rasch analysis requires a specific program to undertake the analysis whereas CTT can be done by hand if required. This is a restriction that assessors cannot overcome, and simply means that if a Rasch analysis program cannot be accessed for any reason then it is impossible to utilise that methodology, and thus in this circumstance CTT is the only viable option for assessment analysis.

As mentioned previously, the desire for future assessment analysis needs to be considered as that may influence the choice of methodologies. For example, if the assessment results are hoped to be compared from year to year Rasch analysis is the most appropriate methodology for that. However, if the analysis is simply being performed to evaluate the assessment knowing that many of them are intended to be replaced in future assessments then CTT may be the better methodology. There is also the potential that if the same assessment is planned to be used repeatedly for a reasonable period of time then the first time the assessment is used it is thoroughly analysed using Rasch analysis and then in future analyses CTT is used to ensure that there are no unexpected and significant shifts in the performance of the assessment tasks and items. Whatever the case, it is important that assessment analysis is conducted regularly, and as such it is reasonable to consider how it is hoped to be conducted in the future and what that means for the analysis currently being undertaken.

5.4.2 Using Analysis to Improve Assessments

Once the considerations around the most appropriate methodologies have been explored, and the types of results that the methodology will provide are known, the next step is using those results to inform the future use of the assessment tasks. There is no information that student analysis can provide that will improve the assessment and item functionality, as these are independent of the student cohort. The measures of the assessment task itself need to be addressed first, as this will inform if there are serious issues across the assessment task. If issues are identified through the analysis of the assessment task there is no way of improving the task based on these measures alone; however, it does provide insight into some of the potential issues that can be expected within the individual item analysis.

When analysing items, any item that is found to be problematic then needs to be replaced or adjusted to resolve the issue. The problem with removing an item and replacing it with something that assesses the same or a similar concept is that unless that item has been used previously there is no way of knowing how that item will function within the assessment. This means that there is the potential that the replacement item may be worse than the item it replaced. As a result of this, in many cases it is more beneficial to improve the problematic items, as not only does this ensure that item content remains constant within the assessment task but it also minimises the risk of further damaging the validity of the assessment. Therefore, it is important that problematic items are evaluated on an item by item basis, as sometimes it is worthwhile spending the time to improve the item (for the reasons just stated), but other times it may be better to remove the item and start over with a new one. This decision can be made based on the results of the analysis, or it may be decided after attempting to improve the item, based on how successful that process was.

Which threshold values are surpassed within CTT analysis can be used to logically determine how the item is causing issues. For example, if an item is classified as too difficult (i.e. $P < 0.30$) but has a high discrimination (i.e. $D > 0.30$) and correlation value (i.e. $r_{bpi} > 0.20$) then the item is likely only well understood by higher ability students. Depending upon the purpose of the assessment task it may be reasonable to keep such an item, or it may need to be simplified. However, if there was an item that had a found difficult by the student cohort (i.e. $P < 0.30$) but a very low discrimination (i.e. $D < 0.30$) and correlation (i.e. $r_{pbi} < 0.20$) that would suggest that lower ability students are having more success on that item than higher ability students. This may be due to factors such as guessing and "test-wiseness", or due to the construction of the item that may be causing higher ability students to select a different option. It can be difficult to evaluate which of these possibilities is the root cause of the issue without further information, and therefore if the issue cannot be determined through a breakdown in the item construction it may simply be less time consuming to replace such an item due to the lack of additional information CTT provides. Similar logical reasoning can be made for any item that surpasses any of the thresholds; however, one of the issues with CTT is that if an item only surpasses one threshold and only does so to a small extent it becomes very hard to rationalise that outcome using the information provided. When this is the case, surpassing the threshold may be treated either as variation within the results of the item, or the construction of the item can be analysed to attempt to rationalise the outcome observed. An alternative approach is that the item can be highlighted, and its performance monitored in future assessments to ensure that it is not a consistent issue. A very common example of items lying outside of thresholds within reason is MCQ items that assess recall and recognition of simple but important facts and concepts, and thus it should be expected that at least some of those items are answered correctly by the majority of the students. This may therefore mean that a difficulty level that surpasses the regular threshold (i.e. above a 90% answer rate) is not uncommon for some of those items; however, that does not make

these items problematic as they are fulfilling their purpose within the assessment. Therefore, even though some items may pass the recommended threshold values it is important to consider what that implies about the performance of the item and if that matches the purpose of the assessment.

Rasch analysis is performed in a similar manner; however, the purpose of the assessment does not need to be taken into consideration when undertaking Rasch analysis as the item difficulty measure informs the expectations that the model places upon it. This means that an item will not be identified as problematic simply because a large proportion of the student cohort answered it correctly, but rather, if the item difficulty is found to be very low, a large proportion of correct answers will be expected by the model. This does not mean that the purpose of the assessment is not a consideration within item analysis, but rather it is approached in a different manner as tools such as the Wright map can be used to compare the item difficulty and student ability measures to ensure that the assessment is adequately targeting the level intended. Interpreting the results of Rasch analysis involves monitoring the deviations from expected values to inform potential problems within the item. For example, if the high difficulty and low discrimination item example from CTT (i.e. $P < 0.30$; $D < 0.30$; $r_{pbi} < 0.20$) is considered, Rasch analysis would inform where the deviations from the expectations are occurring. This means that Rasch analysis can identify if the result is caused by low ability students overperforming or high ability students underperforming and based on that a better decision as to how the item can be improved can be made. This can be achieved using an item characteristic curve (ICC) or using the infit and outfit values which will be able to inform where any significant deviation is occurring.

The approach to improving assessment items through analytical methods is different depending on the methodology used, as Rasch analysis provides more direction to the cause of the issue and thus provides a clearer area to focus on. As discussed previously (see Section 3.3.4) improving an item can only be done through the item stem or the item options, and only one of these can be informed by these methodologies. Breaking down the item options using CTT requires how frequently each option was selected, which can be used to generate the average raw score of the students that selected each option. This can be used to determine which options may be selected more frequently than expected based on the average ability selecting that option, and what options are not being selected. Rasch analysis provides very similar information within its distractor analysis including the option selection frequency, it also gives the average ability of the students that selected each option, which can be used to identify options that are favoured by higher ability students. The reason that Rasch distractor analysis is more reliable than CTT distractor analysis despite being able to generate similar result is due to the dependent nature of CTT compared to the independent measures within Rasch analysis. This provides more confidence that the results seen within Rasch analysis are not being influenced by any outside factors, whereas the potential relationship between item measure and the student cohort within CTT means that there is the potential that this may influence the results seen. By using these analytical methodologies it is possible to not only identify which items are potential concerns for the validity of the assessment, but also obtain some direction as to how the item needs to be changed to ensure it is not compromising the purpose of the assessment.

Using the results provided by CTT and Rasch methodologies is not the only way to analyse assessments and improve upon them. Assessment tasks and items can be reviewed without any mathematical analysis to ensure that the construction of the items matches the purpose of the assessment. However, the breakdown of item construction is most appropriate when it is paired with results provided by an analytical methodology, as not only does this identify the problematic items but it also provide direction as to the areas that may need to be improved. Knowing that a

distractor option is causing issues within the item only identifies the issue, it does not solve it, and thus the use of either methodology will only be able to inform a certain amount before improvements need to be suggested without its guidance. Knowing why an item or distractor is not functioning as expected is still extremely valuable when making changes to improve the performance of an item, as it means that logical reasoning can be used to inform the changes that need to be made rather than changing what appears to be the most problematic aspect and assuming this will improve item performance.

The other approach that can be used to analyse assessments is comparing their performance over time, which can be used to ensure that an assessment is performing consistently and to judge how changes outside of assessment are influencing student performance. If the performance of an assessment changes between uses this could possibly indicate a change in the course has influenced how the students perform within assessments. If CTT is used then there is the potential that any changes observed are due to changes within the student cohort, and while analysis can be performed to identify if that is the case it will always be an influential factor. However, if Rasch analysis is used then the independence of the measures makes it easy to identify the root cause of any changes. If the items are performing significantly differently between assessments then there is a concern that there may be some underlying issues within the items that have not been resolved previously. This may not be limited to significant deviations from the Rasch model but may be due to differential item functioning (DIF) that changes the performance of the item based upon the student cohort. If the students are performing significantly differently between years it may indicate that changes to the course are having an influence on the students' results, either positively or negatively, as based on this research a significant change in the student cohort ability should not be expected. Using this, there is the potential that assessment analysis can be used to track how changes to the course influence the student outcomes, and therefore be used to inform a decision about whether those changes fulfilled their purpose or not.

5.5 Conclusions

Analysing assessment tasks and items is important for ensuring their validity; however, the methodology selected has a significant influence on the capabilities of that analysis. That is why this chapter focused on addressing the question and the objective related to determining approachable and effective methods of MCQ assessment analysis.

What is the most appropriate way to analyse MCQs in order to provide an approachable methodology that can be used to improve assessments?

To identify the most approachable and effective methods to analyse MCQs, and develop a process that can be used to improve any MCQ assessment

The methodologies used in this research were CTT and Rasch analysis, as these two methodologies arguably cover the two opposite ends of the spectrum: CTT is very approachable but only gives minimal information, while Rasch analysis is more involved but gives a wealth of information. In terms of the information that they provide Rasch analysis will always outperform CTT, as it produces the same results and much more, plus the independence of the student ability and item difficulty measures enables alternative analysis approaches. When using CTT there is always a requirement that the identity of the influencing factor has to be justified, as the fact that the student cohort and

the difficulty of the items are dependent upon each other and thus both can influence the results of an assessment. In comparison, Rasch analysis provides a clear and more detailed breakdown of the assessment task, the items within it, and the performance of the students; however, the 'big picture' results are comparable between both methodologies.

The underlying principle of the confirmatory nature of Rasch analysis is responsible for some of the large differences seen when applying the two different methods. This allows for the generation of independent item difficulty and student ability measures and it also means that the expectations of the item are predetermined and thus the assessors cannot form their own expectations for what is reasonable for the item. It could be argued that this is to the detriment of the analysis as there is the potential that the assessment results are not expected to fit the Rasch model, whereas CTT allows for flexibility in the expectations that assessors may have of each item; however, as CTT generates values that are dependent on the student cohort being assessed it means that those expectations are reflective of the expectations of the student cohort and not necessarily of the items themselves. The use of a confirmatory method does allow for deviations away from the expected outcomes to be identified more easily, as it can be incorporated within the analysis itself, whereas the continued consideration of what is expected of an item and how that is reflected within the analytical measures is a time-consuming process.

It is important that the purpose of the analysis is well defined, as this will influence the most appropriate methodology. In the case where only the largest issues within the assessment tasks and items are of concern, CTT is the more approachable method of obtaining results to fulfil that purpose. However, if more information is required to either analyse the performance of individual students or to further break down the results of the task and items, then a more detailed methodology such as Rasch analysis is required. In general, it would be recommended that for assessments that have never been analysed before, a methodology such as Rasch is utilised to ensure that any problems with the assessments can be identified and the reasons for them resolved. After the initial analysis it is then possible to use a simpler methodology such as CTT on every successive use of that assessment, as only large shifts in performance, identifiable by CTT should be of concern. Therefore, while in an ideal world every assessment would be analysed in as much detail as possible using a method such as Rasch analysis, this will not always be the most effective way to achieve the outcomes required of every assessment analysis. Thus, the significance of factors such as time, methodology availability, future assessment plans, and importance of student analysis need to be considered when deciding which methodology is most appropriate and the information that is required to ensure the validity of the assessment.

Chapter 6: Conclusions

6.1 Breaking Down Assessment Tasks and Items

6.1.1 Performance and Consistency of the Multiple-Choice Assessment Format

The multiple-choice question (MCQ) format can be used to fulfil a wide range of purposes, and hence there can be great diversity between two different MCQ assessment tasks in terms of the content and the order of thinking being assessed. There are limitations on what can be assessed by the format, as it has no way to assess student evaluation and creativity; however, outside of those limitations it is possible to construct an item that is able to assess any other aspect of student learning. Like any assessment format the effectiveness of the assessment task is dependent upon its construction, and thus concerns about the format tend to be reflective of issues within item construction rather than the format itself. If the items are constructed to match the purpose of the assessment task, there should be no issue with a MCQ assessment fulfilling its purpose.

While there are concerns about a format that has the potential for statistical anomalies to occur within its results due to factors such as student guessing or 'gaming' the assessment, these should not be considered influential factors within the majority of MCQ assessment tasks. This is because given enough items the chances of a student correctly answering all the items in this way becomes increasingly smaller, and in addition to this it is possible to apply methods to prevent student guessing from influencing an assessment. This can be done either by removing a factor from the students' results based on the number of incorrect answers they provide, or by including negative marking. As MCQs are the quickest formats for students to answer individual items more items can be included within a timed assessment than any other format will allow for, which not only reduces the impact of guessing, but it also means that a wider breadth of content may be included within an assessment task. This also influences how students revise, as they must revise the entire course content. Whilst the counter argument to that is MCQs assess the content at a shallower level than other assessment formats, this is only true for the highest orders of thinking, and thus assessors need to ensure that their items are constructed for the level they wish to be assessing rather than assuming the limitations of the format.

Throughout this research it was consistently seen that there is nothing about the MCQ format that inherently influences the results of the students. Whilst there is the potential that some students may prefer this format over other formats that does not make the use of this format any less valid than any other assessment format. The most important consideration is that the most appropriate assessment format is selected for the purpose of the assessment task. The MCQ format will not always be the most appropriate format, but it is important that it is considered for both its merits and its flaws rather than assuming the limitations of the format before it is used.

6.1.2 Performance of Individual Items

Within this research one of the major questions and objectives was ensuring that the MCQ assessment tasks used within first-year chemistry courses at The University of Adelaide are performing as expected, and that they measure the ability of the students in the content being assessed. This is important as it validates the use of the assessment tasks as tools for evaluating the learning of the students, and it ensures that this research is based on assessment tasks that fulfil their purpose. To address this a simple question and objective were proposed.

Research Question 1:

Are the MCQ items used both previously and currently at The University of Adelaide in first-year Chemistry courses performing as they are expected to?

Research Objective 1:

To assess the items used in MCQ assessments both currently and previously at the University of Adelaide in first-year Chemistry courses to determine whether those items are providing the assessors with information that reflects the ability of the students

This was addressed within Section 3.3, and while it was determined that the assessment tasks were a valid way of assessing the students' ability there were some individual items identified whose results were not reflective of the students' ability. Two different methodologies were applied to the assessment tasks to analyse the entire task and the individual items, but these methodologies gave conflicting results on which individual items are cause for concern. The four courses analysed over a period of four years had a combined total of 261 unique items, 12 of which were identified as problematic items using Classical Test Theory (CTT) (see Appendix 7.8), while 83 problematic items were identified using Rasch analysis (see Appendix 7.9). These large differences can be rationalised based on differences between the two methodologies, such as the parameters used to deem when an item is problematic, the ability to change thresholds that define a problematic item based on the expectations based on that item, and the nuance of the information provided (see Section 5.2.3 for further explanation); however, these large differences change how the performance of the assessment tasks are viewed. Based on the results of CTT only 4.6% of the assessment items are a concern for their ability to measure student learning, whereas 31.8% of the assessment items are a concern based on the results of Rasch analysis. Further analysis of the items showed that only 4 items identified by CTT as problematic reflect major issues, 4 items require more analysis to determine if they are major or minor issues, and 4 items were only minor issues. Using Rasch analysis 41 items that were identified as problematic were considered major issues, 9 may be major but require further analysis, and 33 items were only minor issues. Even though items with major issues likely influenced the results of some of the students undertaking the assessment tasks, the fact that they were spread across multiple different assessments that took place over multiple years means that these items did make the results of any of the assessment tasks invalid.

Whilst there were many items that were found to be problematic, only some of them will have negatively influenced the students' results, as some items may be dysfunctional due to them increasing the results of the student cohort if they are susceptible to answer strategies. An argument can be made that if an item is susceptible to answer strategies then there is no advantage that any one student can obtain from dysfunctional items that is not available to other students. There is also the argument that item issues that are causing students of all abilities to have difficulty answering the item correctly are fair as they effect all students equally. While these theories may be true, there is no reason to risk the validity of an assessment task on the basis that as all the students are required to answer the same items they therefore have the same advantages and disadvantages afforded by those items. It is unlikely that these theories are true across the entire student cohort, as, for example, higher ability students are less likely to be confused by the concepts within an item, and thus poor item construction is more likely to influence their results than a student who is less familiar with the content being assessed as they may not understand the concepts at all. Similarly, the application of answer strategies requires previous experience within assessment tasks and applying those strategies, a requirement that is completely unrelated to the students' ability within the content being assessed, and thus there is potential that some students may apply those skills

more effectively than others. This means that all the problematic items were treated as a risk to the assessment validity when they were identified, and no items were dismissed because all the students were presented with the same items.

The risk of leaving items that were identified as problematic within assessment tasks without addressing them is that they may undermine the purpose of the task, and as such the assessment task may not be reflective of the students' ability. Many of these concerns surrounding problematic items relate to influences that do not correlate to the students' ability, such as the application of answer strategies or the potential for misinterpretation within the items. These issues can be hard to identify before the items are used within an assessment, as despite the best of intentions in how an item is constructed the way that students' approach and react to that item is somewhat unpredictable. Thus, the only way to fix these items is to evaluate them after the students have had the opportunity to answer them, and use the information obtained from that evaluation to address any concerns identified. There are some issues that can be identified before the items are utilised within an assessment task, as some issues can be discovered by reviewing the items construction, ideally in conjunction with a third party. Each individual item identified through the analysis undertaken in this research was broken down in terms of how it was constructed and by any potential issues it contained in the hopes that changes could be made to resolve the problems within the item.

An important aspect that also needs to be considered is items that are not problematic but still inhibit the purpose of the assessment task, as these items provide no additional information about the students. These items typically are either too easy or too hard for the student cohort, or they split the cohort into two halves and provide no information about how the students in each half are different from each other. If an item cannot provide information about the extent of the students' learning relative to other students then, unless that item was specifically included to fulfil a purpose, the item is behaving in a dysfunctional way for that assessment task. This was kept in mind during the process of analysing the items; however, as the identification of problematic items was the main priority rather than optimising the information obtained by each item, for the most part items that provided minimal information were seen as non-issues within assessment tasks.

There is the potential that some of the items identified to be problematic may be fulfilling a specific role within the assessment task, and thus this was considered when evaluating any items identified as being problematic. The simplest example of this is any items that are assessing the students on the fundamentals of the content, as it should be expected that an item testing such content is answered relatively easily by the majority of the student cohort. That does not mean that such an item is problematic despite the poor item fit measures that it may exhibit, as it is important that students are assessed on the entire breadth of their course knowledge, not just the more difficult content. There were occasions within this research where items were reported as being problematic, but upon review they were functioning as expected, which highlights the importance of item expectations in the analysis process.

Based on all the results of the assessment tasks, individual items, and the considerations relating to them it was reasonable to conclude that the assessment tasks analysed within this research were fulfilling their purpose; however, there were individual items that jeopardised the validity of the assessment results. Due to the relatively small number of items that were a significant concern (4 major item issues identified by CTT [see Section 3.3.1 for discussion], 41 major item issues identified using Rasch analysis out of 261 unique MCQ items [see Section 3.3.3 for discussion]) it meant that no

single assessment task contained enough problematic items that they influenced the results provided by the other items. All the items identified were evaluated, as despite the assessment tasks being considered fair evaluators of student ability, improving these items ensures that student results cannot be influenced by the items. The items identified as problematic were also made note of for the remainder of this research, as in some cases it was unreasonable to consider items that were identified as problematic in other comparisons that were made.

6.1.3 Construction of Items

The construction of each item was categorised when they were evaluated to uncover potential reasons for any issues identified. This involves ensuring that the item is constructed in a way that minimises the potential for student confusion, whilst also giving the students the context and content required to answer the item without introducing any way for the students to 'game' the item. This process was expanded upon in this research, as described within Section 3.4.6, to classify the items based upon their construction in alignment with one of the research objectives.

Research Objective 2:

To analyse the construction of MCQ items utilised at The University of Adelaide in first-year Chemistry courses to develop a method of classification for MCQ items

Using some of the key components described in the construction of items, as well as learning taxonomies and accounting for construction aspects unique to MCQs, seven different categories that can be used to describe the construction of an item were generated. These categories are content, taxonomy, type, presentation, process, complexity, and potential item issues. Within each category are different subcategories and factors that are used to describe that item (see Section 3.4.6).

In order to apply this process to other assessment tasks there will be a requirement to include new factors that are not described here, and there is the potential for any of the factors listed to be removed if they are irrelevant to that assessment. Using this process, a description of each item was generated that was used to ensure that the item fits within the purpose of the assessment. Alternatively, this process could be used to evaluate if an assessment task covers the range of items that is required to fulfil its purpose. Within this research this was used not only to categorise the types of items that appeared within the assessment tasks, but also in an attempt to identify if there were any trends within the items that appeared as problematic or showed gender bias. Through this process there were no continuous trends that were observed, but it does give the opportunity to ensure that any item that may be used to replace a problematic item is constructed in a similar manner to the item being replaced. Despite there being no pattern identified using the item categorisation process, it is expected that this process has the potential to be a versatile tool that can be used in conjunction with other forms of assessment evaluation either before or after the assessment has taken place to ensure that assessors are aware of the types of items present within the assessment task.

6.2 Comparing the Performance of Students using Assessment Tasks

6.2.1 Measurement of Student Ability

The purpose of most assessment tasks is to measure the knowledge of the students on the content being assessed, regardless of whether that measure is used to determine student learning or to rank the students based on their ability. This means that it is critical that the assessment tasks are a valid

way to measure the ability of the students, which is why it is important that before the results of the students are further evaluated the assessment tasks themselves and the items used within them are analysed and any issues addressed. If the assessment task is shown to be a valid way of measuring the ability of the students then it is possible to rationalise a detailed breakdown of the students, as was done within this research.

The results of the students should always be analysed with a specific purpose or goal in mind, as it is possible to generate a wealth of information about the students, but without purpose that information is meaningless. It is likely that a statistically significant result about the students will be identified if the results are analysed based on every possible factor; however, this may be due to the amount of data being evaluated and hence it is not representative of a truly significant result. Therefore, it is recommended that if assessors have no discernible purpose for evaluating the results of the students in greater detail, then no such analysis should be undertaken and instead focus should be placed on evaluating the assessment task and items. Within this research the evaluation of the students was focused on the student cohort and not the individuals.

It is also important to remember that even though all this research is based on MCQ assessment tasks and their items, there are other assessment formats that can be used to measure the ability of the students. It is possible to analyse the results of other assessment formats in the same way that was done in this research; however, there are additional considerations that need to be accounted for when doing such an analysis. The most critical of these considerations is the objective nature of MCQ assessment tasks compared to the slightly potentially less objective nature of other assessment formats when a marker is required to evaluate the answers that students provide.

6.2.2 Comparison of Gender Cohorts

The first student comparison that was made within this research was comparing the results of male and female students within Section 3.4 to address one of this research's questions and objectives.

Research Question 2:

Is there a significant difference in the performance of male and female students within MCQ assessments? If so, how can this be addressed?

Research Objective 3:

To compare the performance of male and female students in first year Chemistry MCQ assessments at The University of Adelaide to ensure that any difference in performance is a result of a difference in ability and not due to factors within individual items that influence student performance based on their gender

It is important that if the individual items are being compared to determine if there was a gender bias present within the item that the assumption that male and female students had equal probability of giving the correct answer was true. This meant that the male and female student cohorts needed to be compared to ensure that no differences in performance were expected based on the students themselves. This comparison was made using both CTT (comparing the raw scores of male and female students) and Rasch analysis (comparing the ability measures of male and female students) through the use of independent sample t-tests to determine if the assumption was justified. It was found that in most cases this assumption was true; however, when there was a difference it favoured male students on all but one occasion (see Appendix 7.10 for CTT results and Appendix 7.13 for Rasch analysis results). There was also a difference in the results observed by CTT

and Rasch analysis, which can be rationalised by the independence of the measures produced by Rasch analysis. More confidence was placed in the results of Rasch analysis due to the independence of the measures, as this means that the student comparison could not be influenced by the students' results on items that were potentially gender biased. Within this research, knowing that there was the potential that the ability of the male students was higher within some of the assessments did not change the analysis that was performed. Instead, consideration was placed into whether any statistically significant differences observed may be influenced by the student cohort undertaking the assessment task that were known be have significantly differing abilities. This was harder to rationalise within CTT than Rasch analysis, as Rasch also gives independent measures of item difficulty and thus the statistically significant difference in the student cohort is not expected to impact the item difficulties generated.

The comparison of the individual items (Section 3.4.3 and Section 3.4.4) revealed that several of them did contain statistically significant gender bias (14 items identified by Rasch analysis and 27 items identified by CTT) (see Appendix 7.12 for CTT results, and Appendix 7.15 for Rasch analysis results). The differences between the items identified by both methodologies can be explained by Rasch analysis's ability to generate independent measures, and the differences in how significance is determined by the two methodologies. In the same way that more confidence was placed in the results of the Rasch student comparison due to the independent measures, there is more confidence that the items identified by Rasch analysis have underlying issues.

Most of the gender bias was observed in the courses Chemistry IA and Chemistry IB, with 13 assessment tasks showing statistically significant differences between the male and female student cohorts using CTT (all favouring males), and 12 assessments using Rasch analysis (all favouring males). These courses also contained most of the gender biased items with 21 items identified using CTT (17 male bias, 2 female bias, 2 swapped gender bias) and 10 items identified using Rasch analysis (6 male bias, 4 female bias). There was less gender bias observed within the Foundation courses (Foundations of Chemistry IA and Foundations of Chemistry IB) with only 2 assessment tasks seen having a statistically significant difference in the ability of male and female students using CTT (1 male bias, 1 female bias) and 3 assessment tasks using Rasch analysis (all favouring males). This corresponded to fewer gender biased items within those courses too, with CTT identifying six items (2 male bias, 2 female bias, 2 swapped gender bias) and Rasch analysis identifying four items (1 male bias, 2 female bias, 1 swapped gender bias). The smaller number of statistically significant assessment tasks between male and female students within Foundations courses does indicate that there is potentially a shift in the types of male and female students that enrol within Foundations courses compared to the IA and IB courses, which could be caused by student self-selection. Despite this all of the courses analysed have gender biased items present within them, potentially caused by the construction of the individual items or the specific content being assessed; these need to be addressed to improve the overall validity of the assessment tasks concerned.

The most consistently identified category for gender biased items was the involvement of some amount of logical reasoning, which means that the students were required to take a piece of information given within the stem and determine which option represented the appropriate outcome. Typically, these sorts of items assess the students' understanding of a topic and their ability to apply concepts in new and unique ways; however, not all of the items that required the students to apply logical reasoning were found to be gender biased. In fact, the majority of the logical reasoning items had no problems associated with them whatsoever, and thus despite this being the most common classification of gender biased items it gives no real information about the

potential underlying cause for the issues being observed. Overall, there were relatively few items identified to exhibit gender bias, which made it harder to observe common trends within all of them, as every trend within gender biased items was matched by several items that had the same trend but did not display gender bias. This suggests that multiple factors within individual items are causing the observed gender bias, which may relate to the process that the students are required to follow and the content that is being assessed. There is also the potential that one gender may be better or worse at specific content than the other, and this issue cannot be addressed within the item itself but rather needs to be approached within the course and its teaching pedagogy.

The fact that there is no specific and consistent issue that is causing gender bias also makes these items hard to resolve, as the underlying cause of the issue is not clear. Typically, the easiest way to identify the cause of the gender bias is by inspecting which options are favoured by each gender, as this may be an indicator of what misconceptions are being held by one of the gender cohorts. However, if there are not one or two obvious distractors that are being selected more often by one gender then there is very little information that can be used to inform the changes required. In these instances, the best approach is to improve the items in any way possible by thoroughly reviewing them and then monitoring their performance in future assessments. These items could also be removed from the assessment task and replaced with new items; however, there is no guarantee that the new items will perform any better and they may cause new issues to appear within the assessment task.

The issues observed within the MCQ assessment tasks should not be expected to be limited to MCQ assessments, as in many instances the root cause of the issue could not be accurately determined. If the MCQ format was flawed then it should be expected that there would be a more observable and consistent difference in the performance of male and female students, as well as a larger percentage of items displaying statistically significant gender bias. Therefore, it is reasonable to assume that gender biased items are a potential concern for all assessment formats. However, since so little is known about the underlying causes of gender bias, it is unreasonable to expect that it is possible to remove gender bias from a new assessment task. Thus, it is important that gender bias is a consideration when reviewing the performance of an assessment task and its items to ensure that there is none present that could be influencing the results of the students.

6.2.3 Changes in Test-Retest Student Performance

Knowing that the problematic items within the assessment tasks did not invalidate their ability to measure the ability of the students due to the spread of those items over multiple assessment tasks and courses meant that it was possible to compare the performance of students at different time intervals, completing another research objective.

Research Question 3:

Do students show differences in their performance in MCQ assessments at different points in a semester? If so, how?

Research Objective 4:

To compare item and student performance within first year Chemistry assessments over the period of a semester, across multiple years, and against Biology courses using MCQ assessments undertaken at The University of Adelaide to determine if there are any differences in performance, and if they these changes are a result of the items or the students

A test-retest methodology is used in the courses analysed within this research to give the students the opportunity to show improvements in their learning; however, it also provides an opportunity for students who were unable to attend one of the assessment tasks the chance to sit that assessment task. It is important from both an assessment validity and student comparison standpoint that shared content is used within both assessment tasks. It is also important for assessment validity that both tasks are equivalent, and thus the results of both assessment tasks assess the students to the same level. When comparing the results of the assessments there needs to be items that are utilised in both assessment tasks to ensure that it is valid to compare between the two assessments, as CTT can only compare the results of identical items and Rasch analysis requires at least five overlapping items to be able to link multiple assessment tasks.

The comparison of student results across the two assessment tasks was made using both CTT and Rasch analysis in Section 4.2, where the items were first compared to ensure that they were performing similarly in both assessments before the results of the students were compared. However, as CTT can only compare the results of identical items there is a concern that the assessment validity may be affected by removing the results of items that are not shared between the two assessment tasks. The requirement for shared items meant that it was not possible to compare the results of any of the assessment tasks within the Foundations courses, as they had a maximum of three items that were shared between the assessment tasks, which is not enough for either methodology. Despite these methodological differences the results of both methods are very similar, as they both show a statistically significant improvement in the mean result of the student cohort within the second assessment. However, even though every comparison showed a statistically significant increase in the mean result of the student cohort, not every individual student improved, and the average increase and the average decrease in the students' results was similar. This means that the only reason that statistical significance was observed was due to a larger proportion of the students improving.

The students that only undertook one of the two assessment tasks were also analysed, and it was observed that generally the students who only undertook the assessment during the semester were amongst the highest ability students, and the students who only took the assessment in the redeemable section of the final exam were amongst the lowest ability students. This observation suggests that student attitudes had a significant influence on their results. For example, it is possible that students who only undertook the assessment in the redeemable section of the final exam they may have done so because it was coupled with the final exam, and thus they were required to be present and prepared for an assessment. It is also possible that the students were unable to attend the lecture tests for a legitimate reason; however, it is unreasonable to assume that this is true for all the students who only undertook the assessment within the redeemable section of the final exam. Similarly, any students who sat the lecture tests and achieved a result that they believed to be adequate for their personal goals may not have attempted the second undertaking of the assessment if they did not believe that they could improve upon their result or saw no reason to try and improve upon their result.

The consistency of the statistically significant improvement within the student cohort demonstrates that student learning is not always completed after they have been introduced to the content, but rather each individual may have their own process for retaining and applying that information. Therefore, while constructing assessment tasks with the intention for students to resit the same or a very similar assessment later in the course is not suitable for all assessments, nor all courses, it is

important to consider the timing of assessments and its use as a learning tool when planning and constructing assessment tasks.

6.2.4 Differences in Yearly Cohort Performance

Mentioned within research objective 4 was the goal to compare student cohorts over multiple years to determine if the ability of the students enrolling within courses was changing over time, as stated within one of the research questions.

Research Question 4:

Do student cohorts show differences in performance over multiple years? If so, how?

It is not an uncommon sentiment that student ability changes between years: sometimes it may be believed that the students are improving but usually the sentiment is that in general students are getting worse over time. However, despite that sentiment it is hard to compare the circumstances of one student cohort to another, and thus often the comparison has too many variables that cannot be controlled to allow for justifiable conclusions to be made. The use of MCQ assessments within this research helps alleviate many of those concerns, as it is an objective format and there is a large number of items shared between the years being compared. Whilst there is no way to account for larger changes within the course itself these sorts of changes are rare and are a consideration outside of the analysis.

As described within Section 4.3 comparing the items across multiple years was done first to ensure that the items were performing consistently and thus were not significantly influencing the student cohorts differently. For Rasch analysis there were no issues identified through this comparison; however, using CTT there were multiple occasions where the mean item difficulty was statistically significantly different between years. It was assumed that as the item difficulty measures were dependent upon the student cohort that this indicated a change within the student cohort as opposed to a change within the items, but this result has two consequences. First, it needed to be addressed after the student comparison was made to ensure that the assumption made is reasonable given those results, and second, it lowers the confidence in the results of the CTT student comparison, as the assumptions being made cannot be completely justified.

When the comparison of the student cohorts was made there were very minimal differences observed between the mean ability of the student cohorts from different years, with only two cohorts identified as being statistically significantly different (Chemistry IB 2012 and Foundations of Chemistry IB 2012). CTT only identified the Chemistry IB 2012 cohort to be statistically significantly lower performing than the 2013 and 2014 cohorts, whilst Rasch analysis identified the same result and that the Foundations of Chemistry IB 2012 student cohort performed statistically significantly better than the other student cohorts. When this analysis was performed a compromise that had to be made for CTT was that student percentages across all MCQ assessments were used to compare the students. This was because there is no other way to account for the potential differences in the number of items that the students may have undertaken due to the optional/redeemable assessment tasks. Thus, there is the potential that the results may be skewed by the attitude and approach that the students take to each assessment task; however, as only minimal differences were observed it is reasonable to assume that those approaches and attitudes are somewhat consistent over multiple years. In comparison to this, Rasch analysis does not require the students to undertake the same number of items to generate a comparable student ability measure; however, the student approaches and attitudes still have the potential to influence their ability measure.

The statistically significant differences observed within the CTT item difficulty comparison did not match the statistically significant differences observed within the student cohort, which does suggest that item performance may be changing over time. It is possible that even though the mean results of the student cohorts are not statistically significantly different between most years changes in how cohorts respond to each item could influence the item difficulty comparison; however, without further analysis this cannot be determined. The item difficulty measures generated by Rasch analysis do not show any significant changes between years, and therefore support the theory that the changes in CTT item difficulty is a result of the students rather than the items. However, using a different analytical methodology to justify the results of CTT makes the use of CTT superfluous, which impacts the degree of confidence in the CTT results.

The fact that the mean result of student cohorts did not change significantly between years except on 2 of the 16 different yearly cohorts analysed suggests that there is consistency in both the types of students enrolling within the courses, and the learning achieved throughout the course. It was expected that this would be true for Chemistry IA and IB, due to having prerequisites, but there was the potential that the Foundations courses would experience more variance due to them being open to any student regardless of their previous experience. Even though Foundations courses did see slightly more variance within their mean results, the fact that only one student cohort was significantly different (the same number that was seen within Chemistry IA and IB) means that despite there being no barriers to entry for the course, the same type of students seem to enrol every year.

If the expectation becomes that student cohorts consistently perform to the same ability level each year this can be used to determine how changes to the course may be influencing the student cohort. These changes could either represent minor changes, such as the way the content is delivered, or larger changes, such as the topics within the course. In each instance, before a comparison is made it is important to consider whether the comparison is justified based on the extent of the changes that were made to the course. This needs to be considered based on the assessment tasks (i.e. is there enough commonality within the assessment tasks to justify the comparison), and also considering changes to the course structure and learning objectives to ensure that what is expected from the students is consistent in both years being compared. This is why it is unreasonable to compare student cohorts who undertook the course during vastly different time periods, as both the expectations that were had of the students and the objectives of the course change over time, disregarding the fact that the assessment tasks would also be drastically different. If such a comparison can be justified, despite any changes made to the course, the analysis will be able to inform whether the student cohorts performed significantly differently from each other, and thus whether that change positively or negatively impacted the mean result of the student cohort. Based on those results and the purpose of the change an informed decision can be made as to whether the change had the intended effect and if it should be kept within the course.

6.2.5 Comparing Performance Across Courses

The last aspect of research objective 4 was the comparison of student ability across different courses from different disciplinary areas to determine if there is consistency in the relative ability of the students across those courses, which is one of the research questions.

Research Question 5:

Is it possible to compare student results across multiple courses from different disciplinary areas? If so, do students show similar performance across multiple courses?

The common sentiment is that students who are amongst the highest achieving students in one course are likely to be amongst the highest achieving students within their other courses. However, whilst there are several potential reasons that students may show a consistent level of ability compared to their peers there are also several reasons why this may not be true. If a student shows a consistent level of ability across multiple courses it would to be reasonable to attribute this to the student's approach and attitude towards coursework, as well as some level of ability to learning content. Conversely, there is no reason that students' attitudes and approaches to courses are consistent, as a student's motivations and expectations may differ between courses, and the student's abilities may be better suited to some courses over others. Hence this comparison is to determine the consistency of students over multiple courses, which could potentially be used to identify traits that lead to academic success and demonstrate the relative level of the courses' assessment tasks.

As there are no shared assessment tasks or items between courses it meant that the only way to link the results of the courses was through the students who undertook both courses. In theory, linking the assessment tasks using students would generate comparable item difficulty and student ability measures. As the assessment tasks needed to be linked through shared students it meant that it was impossible to carry out this analysis using CTT. Before the courses were attempted to be linked it is important that the comparison is justified, as if the two courses assess vastly different abilities then there is no reason to believe that the students will perform in the same way. Within this research chemistry and biology courses were attempted to be compared as it was believed that it is reasonable that these two courses require similar abilities from the students.

The issue that was identified within this research (described within Section 4.4.2) was that there were too many assumptions that needed to be made regarding Rasch analysis of the students' ability that could not be justified. This resulted in there being no valid way to identify students that could be used as anchors between the two assessment tasks, as it could not be confirmed whether those students were performing differently between the courses until the assessment tasks were compared. As an alternative to anchoring the students, the results of the assessment tasks were racked and analysed. This assumes that student ability is consistent across both assessments, which was tested using the dimensionality measure. Based on the dimensionality results of this analysis it suggests that there are multiple significant influences on the measures, which therefore means that the assumption of unidimensionality is not justified within the analysis and thus the results may be flawed. This indicates that the comparison between these two courses cannot be made because the underlying ability that is being assessed by the two courses is different, and hence the comparison is not justified.

Theoretically, this comparison is possible if two courses were identified that were expected to share the same student ability (e.g. Chemistry IA and Chemistry IB, or Chemistry IA and a second year undergraduate Chemistry course); however, to undertake this comparison there needs to be enough shared items to link the assessment tasks. Thus, unless two courses already have shared assessment items this comparison needs to be premeditated to ensure that items can be constructed that are relevant to both courses while still providing information about the students' learning within each course. Before that approach is taken there should be a predefined purpose for this comparison, as

without a purpose there is the potential that the assessment validity could be compromised by the shared items for no reason.

6.3 Methodology for Assessment Analysis

6.3.1 Classical Test Theory

Throughout this research CTT has been shown to be an approachable and efficient methodology for carrying out most of the analysis and comparisons made within this research. However, many of the comparisons made did not use CTT methods but utilised the same assumptions that are made by CTT to apply comparisons using the student raw scores. CTT is a student result focused methodology, and as such all the information that it generates is dependent on the student cohort that is being assessed. Therefore, the information that it provides may not be representative of the assessment task and items, but rather representative of the students and their assessment results. Because of this CTT cannot generate information on the students themselves, as it assumes that the raw score represents the students' true score plus a random error that cannot be accounted for. While this was not a large issue within this research, depending on the purpose of the assessment task there is the potential that this could impede the assessment analysis.

One unique aspect of CTT is the ability to adjust the expected outcomes of the analysis depending upon the expectations placed upon individual assessment tasks or items. This means that while there are recommended thresholds in terms of where each measure is expected to lie it is possible to rationalise any measure produced based on the expectations of what is being analysed. However, this also means that there is the potential that measures that lie outside of the expected thresholds may do so as a results of their function, and hence they may be considered problematic despite performing their purpose. Thus, this means that often the results of CTT needs to be considered with the purpose and expectations of the assessment and the students undertaking it in mind.

After reviewing the measures given by CTT, action needs to be taken based on those results either by improving or removing items that were identified to be problematic. One issue with CTT is that it provides no additional information to the assessors about the origin of any issues within the items. There are ways to obtain information about the performance of each option in an MCQ item; however, this requires additional analysis that is beyond the scope of the usual methods employed by CTT. Therefore, in most cases where CTT is used the improvement of items requires the assessors to act based on their own opinions of where the item may be flawed.

Student comparisons utilising the assumptions of CTT produce results that are comparable to other more detailed methodologies. However, the issue with undertaking these comparisons is that there are often assumptions and compromises that need to be made to justify the comparison. Thus, while there is the potential to undertake assessment comparisons in this way, there will likely be some doubt in the results produced, and therefore undertaking the comparison using a more detailed methodology will provide more confidence in the final results.

CTT is a methodology that excels at ensuring there are no major problems within an assessment task or any of the items it contains. It is not ideal for analysing the assessment task or the individual items in detail and has no capability to generate information about individual students and their performance. This makes it a methodology that is well-suited for assessment tasks that are classified as low stakes, which may be due to them being purely formative assessments, or they may only be a

minor assessment task. This is because CTT will be able to identify the issues that are the greatest concerns for assessment validity, but it will not be able to identify all of the issues that may influence the results of the students. However, if an assessment task is classified as 'low-stakes' it may be considered less important that every potential issue is identified and instead only the largest influences are identified and corrected.

6.3.2 Rasch Analysis

In this research, often more confidence was placed in the results that were produced using the Rasch model due to the information that it provides and the assumptions it can account for. Utilising the Rasch model as a method of analysis was a more time-consuming process to begin with; however, once the model was well understood, the actual application of the model to the assessment tasks was not a lengthy process. An important aspect when applying a more detailed method such as Rasch analysis is knowing what information is desired from the analysis, as otherwise there is the potential to overload on information that is not relevant to the purpose of the analysis.

Rasch analysis is a confirmatory model, and as such it assumes that the results of the assessment task and the performance of the individual students and items will follow the model's predictions. The model itself predicts the probability of a potential outcome occurring (in this research the probability of the students selecting the correct answer) based on the student ability and item difficulty measures. The key aspect that separates this methodology from others is that because of the assumptions of the model it generates measures of student ability and item difficulty that are independent from each other. This is a major reason why more confidence was placed in Rasch analysis, as the student cohort and the assessment items do not influence the measures of each other. However, as the two independent measures are still placed on the same relative scale, it allows the model to make predictions about the expected outcomes and compare those to the observed ones.

The confirmatory nature of the Rasch model could be seen as a negative, as it generates its own expectation of how the assessment tasks, items, and the students are expected to perform. However, the expectations that the model has of these factors is informed based on the measures generated using the assessment results. As the expectations of the Rasch model shift based upon those measures there is no issue with items that may be included to fulfil a specific purpose within the assessment tasks (e.g. assessing the basic concepts of a topic). Therefore, issues that are classified as significant by the model are identified because they do not match the expectations that the model has of them based on their measure.

The information that the model gives as to why an issue is identified as being significantly different from the expectations of the model was used to speculate upon the root cause of the issue. There is the potential that an issue is identified for reasons that are not considered to be a concern by the assessors (e.g. student outliers); however, if the reason is identified then assessors can choose how to proceed based on that information. It is also possible to use the model to analyse individual students to determine if they are deviating from how the model expects them to perform based on their ability. While this was not utilised within this research, this has the potential to identify students that may be cheating or applying answer strategies in the assessment, as their answer pattern may not match the expectations that are had of them based on their ability. There are also other applications for student analysis if more is desired to be known about the students; however, when large cohorts are being analysed, as was the case within this research, it can become easy to

be distracted by all the information generated. Thus, student analysis only needs to be undertaken if it is seen to be relevant to the purpose of the analysis.

The information that the model generates can be used to help improve any items that are identified as problematic by using item option and student selection analysis. It is possible to identify if any distractors are behaving unexpectedly; however, if the options are not the cause of any issues Rasch analysis does not provide information on how the students may be interpreting any other aspect of the items construction. In this instance the assessors would be required to evaluate the item based on its construction and their own opinions. However, even in that instance, the fact that the item options are not the root cause of the issues can be useful in determining the best approach to fixing the item.

Due to the student ability and item difficulty measures being independent from each other, making comparisons over two different time intervals is relatively easy, assuming the assessment tasks being compared can be linked. In most cases, linking the assessments was not a concern due to the large number of shared items within the MCQ assessment tasks being compared, and the comparison of the item difficulty was used to confirm that there was no issue linking the assessment tasks. The independence of the measures meant that there was a high degree of confidence in the results of any student comparison, as there was nothing that could influence the comparison except for the ability of the students themselves.

Rasch analysis provides a large amount of information regarding an assessment task, the individual items, and the students that undertake the assessment. This makes it a methodology that is suited to any analysis, but particularly the analysis of high stakes assessments. It is crucial that in any high stakes assessment tasks, which may influence the students' future, there are no outside influences that could impact the results. Therefore, it is important that every aspect of the assessment task is analysed to ensure that anything that could influence the validity of the task is brought to the attention of the assessors who can then decide how to proceed.

6.3.3 Aiming to Improve Assessments

The purpose of any assessment analysis fundamentally relates to improving the assessment task so that it becomes a more effective tool. However, applying an assessment analysis is not always a straightforward process, which is why a research objective and question focused on this task.

Research Question 6:

What is the most appropriate way to analyse MCQs in order to provide an approachable methodology that can be used to improve assessments?

Research Objective 5:

To identify the most approachable and effective methods to analyse MCQs, and develop a process that can be used to improve any MCQ assessment

The first step in any assessment analysis should be considering what the purpose of the assessment task is, as this will inform decisions regarding the expectations placed upon the items and the students. The purpose of the assessment task can range from ranking student ability to determining student learning throughout a period. Whenever an assessment task is reviewed or changed it should be done with its purpose in mind, and changes need to be made with the aim of making the assessment task more suited to fulfilling its purpose.

Reviewing an assessment task before it has been undertaken by the students should be an important part of an assessment tasks construction, as it is possible that some issues may be identified by reviewing the construction of each item. This ensures that the assessment task is as well constructed as possible before the students undertake it, and as such it minimises the chances that there are influences unrelated to the content being assessed. Reviewing the items after the assessment task has been used follows a similar process, except the items that caused problems within the assessment task will be identified, and thus the review is more specific to those items. Even with the additional information that item analysis provides there are issues within items that will not be able to be identified based on the results of the analysis. Thus, being able to review an assessment task without any additional information is an important skill for assessors to have.

When the assessment is analysed using a specific methodology it is important to consider how this analysis is reflective of the purpose of the assessment. If the purpose of the assessment task is to obtain a rough idea of how the students' learning is progressing, then the methodology required only needs to identify issues that will greatly influence student results. If the purpose of the assessment task is to obtain a highly accurate gauge of the students' learning, then a highly detailed methodology is required to ensure that there are no influences that could affect the results of the assessment task. Therefore, there is never one methodology that is most appropriate in every situation, as assessment tasks have different purposes and different expectations of their outcomes. However, in an ideal world every assessment would be analysed using a more detail orientated methodology, such as Rasch analysis, and the results of the analysis can be interpreted as required.

Assessment tasks should always be reviewed in some capacity to ensure that not only are they a valid way to measure student ability, but also to ensure the assessment task fulfils its purpose. Ideally, the best way to review assessment tasks is by first reviewing them before they are even used, even if this simply involves checking the spelling and grammar of each item. Then before the results of an assessment task are analysed it is important that the purpose of the analysis is clearly defined to ensure that the information most relevant to fulfilling this purpose is used. The next step is to review the assessment task and items' results, which ideally would be done using a methodology that provides the most detail about the assessment; however, depending on the purpose of the assessment task that is not always necessary. The last step in this process is to act on the results of the analysis, which involves either replacing or rewriting specific items within the assessment task using the information that was obtained through the analysis, or fulfilling some other outcome based on the purpose of the analysis.

6.4 Future Directions

6.4.1 Evaluation of Future Assessments

Within this research only results from 2012 – 2015 assessment tasks were used due to the amount of time that was required to prepare the data for analysis whilst maintaining the privacy of the students. Therefore, the most obvious future application of this research is to update the assessment task analysis with the most current MCQ assessments that were undertaken. This would enable any new items introduced to be analysed and evaluated and allow those years to be connected to the comparisons made within this research to see if the trends continue. These trends include items identified as problematic, the differences in student test-retest results, and comparing the yearly cohorts. There is also the possibility to apply what was learnt within this research to the

items currently used within the assessment tasks to address any item construction issues identified and to evaluate the assessment task after they are used to ensure that the research can be successfully applied to repair problematic items.

There is also the potential to expand and analyse other courses across the university that utilise MCQ assessment tasks to assess the students. This can be used evaluate those assessments and ensure that there are no problematic items within them, and to compare the trends within those courses to the trends identified within this research. This is important as the purpose of the assessment task is likely to be different between the courses, which means that the format is likely being applied in different ways. This can be used to ensure that the trends identified within chemistry MCQ assessment tasks apply more broadly across all MCQs, or identify if there are significant differences that need to be addressed to analyse assessment tasks that are used in different courses.

A more unlikely but possible potential application of MCQ assessment analysis is comparing assessment tasks across different educational institutions within the same course topics. This would require a large amount of cooperation and forethought to undertake, as there would need to be an alignment in the content being assessed, the purpose of the assessment task, and a few shared items within the assessment tasks. However, if it is undertaken it could provide a wealth of information about both the students and the assessment tasks in terms of how comparable they are to each other. This could be used to ensure that there is consistency in the level the students are being assessed at across different institutions, as it would be expected that similar outcomes are obtained by the students at multiple institutions. Alternatively, it could simply be used to gauge the differences within the student cohorts enrolling in similar courses at different institutions, which has the potential to provide additional information about the types of students that attend each institution.

6.4.2 Refining Item Breakdown Classification

Within this research a method to categorise the construction of MCQ items was created to determine if there were any patterns to the types of items that were being identified as problematic or behaving abnormally. Whilst it was not successful in identifying any patterns, there is the potential that this categorisation process can be applied more broadly to assessment tasks to ensure that their construction matches the purpose that the assessors designed it for. It could also be used when replacing items within an assessment task, as if the objective is to replace an item with something that assesses similar content in a similar manner then the categorisation of each item can be compared to ensure that they are comparable how they function. There is also the potential that this could be used to identify issues in the diversity of the assessment task, and potential areas that an assessment tasks is not targeting.

To refine the categorisation process such that it could achieve these goals, it needs to be reviewed and applied to a more diverse range of items to ensure that every item can be categorised without compromising the factors assigned to it. Further refining of the process will also help other researchers apply this method, as it can help to accurately define each of the categories present within the process and help minimise any confusion others may experience applying the process to their own assessment tasks. Utilising a diverse range of items requires assessment tasks from different disciplines, and ideally different levels of education. Undertaking the process of item categorisation with that level of assessment diversity should ensure that the process is applicable across all assessment tasks. Applying the process in this way will require changes to be made from

its current iteration, as currently there is a large focus on categories relevant to chemistry assessment tasks that will not be relevant to every assessment task. Thus, there is the potential to create a process that includes numerous categories that are more relevant to some assessment tasks than others and assessors will either be able to choose the categories they believe are most relevant to their assessment task or apply the process with all of the categories for the most comprehensive categorisation of their items.

This process could also be used to examine the performance of individual items and item types in greater detail if the categorisation system is linked to item analysis. By knowing what an individual item has been categorised as it would be possible to consider how the items with those categorisations tend to perform within assessment tasks. Therefore, it may be possible to identify categorisation pairings that tend towards specific outcomes, whether those outcomes relate to the difficulty of the items or how students interpret those items. If this can be determined it allows for a better review of the assessment task before it is undertaken by the students, as more accurate predictions can be made about how the items are expected to perform. This also has the potential to help improve the revision of items after analysis, as it may be possible to link specific issues seen within item analysis to specific categories within the item. A hypothetical example of the potential application would be if it was observed that a significant number of students who were not expected to correctly answer the item did so. Upon reviewing the item using its categorisation it may be known that this is commonly caused by items that contain comprehension and logical reasoning. Therefore, in this hypothetical scenario if those aspects of the item are addressed in some way it should theoretically remove the issue from the item. To be able to apply the process in this way requires more iterations and learning how the categories may influence individual item performance, both of which can be achieved through more item analysis.

Another aspect of this process that can be improved upon is ensuring that how the categorisation of an assessment task or an item relates to the purpose of the assessment is well understood. This would be used to ensure that the construction of the items used within an assessment task match the purpose of that assessment. Whilst the basics of matching items to the purpose of an assessment task does not require a categorisation system (e.g. what is being assessed and the level being assessed), as that can be judged when reviewing the items, finer details can be recognised through the categorisation process. This would involve the identification of ways that item construction work towards the completion of a specific purpose within an assessment task, and thus this knowledge could be used to define the expectations of each item's categorisation within an assessment task.

6.4.3 Potential Indicators of Student Performance

Student performance was only considered within this research as a point of comparison over different time intervals; however, there is the potential that by including more information about the students within the analysis then factors that are indicators of success or risk could be identified. This would involve placing a higher emphasis on the results of the students and the inclusion of factors that may potentially have an influence on their result. Common influences that are worth considering are factors such as the student's age, socioeconomic background, and their previous education experience. These can be used to create distinct student groups and there is the potential that these student groups perform differently within assessment tasks compared to other groups. These factors may affect the attitude, approach, and knowledge that the students have, which are all considered to be influential in the outcomes obtained by students within assessment tasks, and thus the factors themselves may have a significant influence on student results.

Another factor that is worth considering is the student's motivation for the course that they are undertaking, as this could influence the amount of effort that they apply within the course and the expectations they have of their own results. One way to include this as a factor within the analysis is whether the course was undertaken as an elective, or if it was a requirement of their program of study. This has the potential to separate students who are only undertaking the course because they are required to, and the students who are undertaking the course because they want to. More detail can be included by considering the student's program, and to what level they are required to undertake the content. For example, some engineering students at The University of Adelaide are required to undertake first year chemistry courses and some chemistry courses in second year. In comparison to this, science students who are majoring in chemistry will be expected to undertake chemistry in all three years of their degree, and thus it would be expected that the science students are more invested within the course than the engineering students. Therefore, there is the potential that a student's program could be used to predict their attitude and motivation towards a course, and thus be a potential indicator of how that student is expected to perform within the course.

The factors that influence student performance can be used to both identify students who are potentially at risk, or to improve student learning by promoting actions that make other students successful. If, for example, it was found that older students were consistently higher achievers then obviously the students cannot be aged to improve their results, but there is potential that some aspect associated specifically with older students is the key factors that can be encouraged. This may be a specific approach or attitude that these students have, or it may be related to having more experience with assessment tasks. Whatever the case, if the underlying reason for their success can be determined, it could provide actionable information that can be used to help improve student learning and assessment performance. It is also possible that the difference is due to a factor that needs to be discouraged, such as applying 'test-wiseness' to increase assessment results, in which case there is the potential for the assessors to learn and improve their assessment tasks to ensure that this is no longer a significant influencer of student performance within that student group.

References:

1. MacLellan, E. Assessment for Learning: The differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education* 2001, 26, 307-318.
2. Newstead, S. The Purposes of Assessment. *Psychology Learning & Teaching* 2004, 3, 97-101.
3. Cheng, L.; Fox, J. Why Do We Assess? In *Assessment in the Language Classroom: Teachers Supporting Student Learning*, Macmillan Education UK: London, 2017; pp 1-29.
4. Fuentealba, C. The Role of Assessment in the Student Learning Process. *Journal of Veterinary Medical Education* 2011, 38, 157-162.
5. Black, P.; Wiliam, D. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability(formerly: Journal of Personnel Evaluation in Education)* 2009, 21, 5-31.
6. Andrade, H.; Cizek, G. J. *Handbook of formative assessment*. Routledge: 2010.
7. Juwah, C.; Macfarlane-Dick, D.; Matthew, B.; Nicol, D.; Ross, D.; Smith, B. Enhancing student learning through effective formative feedback. *The Higher Education Academy* 2004, 140.
8. Hattie, J.; Timperley, H. The power of feedback. *Review of Educational Research* 2007, 77, 81-112.
9. Snyder, B. R. *The Hidden Curriculum*. Knopf: 1971.
10. Elton, L. R. B.; Laurillard, D. M. Trends in research on student learning. *Studies in Higher Education* 1979, 4, 87-102.
11. Newble, D. I.; Jaeger, K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983, 17, 165-171.
12. Crooks, T. J. The Impact of Classroom Evaluation Practices on Students. *Review of Educational Research* 1988, 58, 438-481.
13. Frederiksen, J. R.; Collins, A. A Systems Approach to Educational Testing. *Educational Researcher* 1989, 18, 27-32.
14. Wall, D.; Alderson, J. C. Does Washback Exist? *Applied Linguistics* 1993, 14, 115-129.
15. McDaniel, M. A.; Blischak, D. M.; Challis, B. The Effects of Test Expectancy on Processing and Memory of Prose. *Contemporary Educational Psychology* 1994, 19, 230-248.
16. Bailey, K. M. Working for Washback: A Review of the Washback Concept in Language Testing. *Language Testing* 1996, 13, 257-279.
17. Van Etten, S.; Freebern, G.; Pressley, M. College students' beliefs about exam preparation. *Contemporary Educational Psychology* 1997, 22, 192-212.
18. Struyven, K.; Dochy, F.; Janssens, S. Students' perceptions about evaluation and assessment in higher education: a review. *Assessment & Evaluation in Higher Education* 2005, 30, 325-341.
19. Al Kadri, H. M.; Al-Moamary, M. S.; Magzoub, M. E.; Roberts, C.; van der Vleuten, C. Students' Perceptions of the Impact of Assessment on Approaches to Learning: A

Comparison Between Two Medical Schools with Similar Curricula. *International Journal of Medical Education* 2011, 2, 44-52.

20. Cilliers, F. J.; Schuwirth, L. W.; Adendorff, H. J.; Herman, N.; van der Vleuten, C. P. The Mechanism of Impact of Summative Assessment on Medical Students' Learning. *Advances in Health Sciences Education : Theory and Practice* 2010, 15, 695-715.

21. Morphew, J. W.; Silva, M.; Herman, G.; West, M. Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering. *Applied Cognitive Psychology* 2019, 34, 168-181.

22. Armbruster, P.; Patel, M.; Johnson, E.; Weiss, M. Active Learning and Student-centered Pedagogy Improve Student Attitudes and Performance in Introductory Biology. *CBE-Life Sciences Education* 2009, 8, 203-213.

23. Sambell, K.; McDowell, L. The construction of the hidden curriculum: messages and meanings in the assessment of student learning. *Assessment & Evaluation in Higher Education* 1998, 23, 391-402.

24. Ertmer, P. A.; Newby, T. J. The expert learner: Strategic, self-regulated, and reflective. *Instructional Science* 1996, 24, 1-24.

25. Chin, C.; Brown, D. E. Learning in Science: A Comparison of Deep and Surface Approaches. *Journal of Research in Science Teaching* 2000, 37, 109-138.

26. Entwistle, N. J.; Peterson, E. R. Conceptions of learning and knowledge in higher education: Relationships with study behaviour and influences of learning environments. *International Journal of Educational Research* 2004, 41, 407-428.

27. İlhan beyaztas, D.; Senemoglu, N. Learning Approaches of Successful Students and Factors Affecting Their Learning Approaches. *Education and Science* 2015, 40, 193-216.

28. Watkins, D.; Dahlin, B.; Ekholm, M. Awareness of the Backwash Effect of Assessment: A phenomenographic Study of the Views of Hong Kong and Swedish Lecturers. *Instructional Science* 2005, 33, 283-309.

29. Chapell, M. S.; Blanding, Z. B.; Silverstein, M. E.; Takahashi, M.; Newman, B.; Gubi, A.; McCann, N. Test Anxiety and Academic Performance in Undergraduate and Graduate Students. *Journal of Educational Psychology* 2005, 97, 268-274.

30. Sarason, I. G. Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology* 1984, 46, 929-938.

31. Cassady, J. C.; Johnson, R. E. Cognitive Test Anxiety and Academic Performance. *Contemporary Educational Psychology* 2002, 27, 270-295.

32. Welsh, P. How first year occupational therapy students rate the degree to which anxiety negatively impacts on their performance in skills assessments: A pilot study at the University of South Australia. *ergo* 2014, 3.

33. Cuff, B. M. *Perceptions of subject difficulty and subject choices: are the two linked, and if so, how?*; Office of Qualifications and Examinations Regulations: London, 2017.

34. Smart, G. The Multiple Choice Examination Paper. *British Journal of Hospital Medicine* 1976, 15, 131.

35. Douglas, M.; Wilson, J.; Ennis, S. Multiple-Choice Question Tests: A Convenient, Flexible and Effective Learning Tool? A Case Study. *Innovations in Education and Teaching International* 2012, 49, 111-121.

36. McCoubrie, P.; McKnight, L. Single Best Answer MCQs: A New Format for the FRCR Part 2a Exam. *Clinical Radiology* 2008, 63, 506-510.

37. Hammond, E. J.; McIndoe, A. K.; Sansome, A. J.; Spargo, P. M. Multiple-Choice Examinations: Adopting an Evidence-Based Approach to Exam Technique. *Anaesthesia* 1998, 53, 1105-1108.

38. Xu, X.; Kauer, S.; Tupy, S. Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology* 2016, 2, 147-158.

39. Schuwirth, L. W. T.; Van Der Vleuten, C. P. M. Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education* 2004, 38, 974-979.

40. Hubbard, J. K.; Potts, M. A.; Couch, B. A. How Question Types Reveal Student Thinking: An Experimental Comparison of Multiple-True-False and Free-Response Formats. *CBE Life Science Education* 2017, 16, 1-13.

41. Lau, M. P. A Theory of Multiple-Choice Examination. *Medical Education* 1972, 6, 61-67.

42. Anderson, J. For Multiple Choice Questions. *Medical Teacher* 1979, 1, 37-42.

43. McCoubrie, P. Improving the Fairness of Multiple-Choice Questions: A Literature Review. *Medical Teacher* 2004, 26, 709-712.

44. Gierl, M. J.; Bulut, O.; Guo, Q.; Zhang, X. Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research* 2017, 87, 1082-1116.

45. Haladyna, T. M.; Downing, S. M.; Rodriguez, M. C. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education* 2002, 15, 309-334.

46. Dell, K. A.; Wantuch, G. A. How-to-guide for writing multiple choice questions for the pharmacy instructor. *Currents in Pharmacy Teaching and Learning* 2017, 9, 137-144.

47. Abozaid, H.; Park, Y. S.; Tekian, A. Peer review improves psychometric characteristics of multiple choice questions. *Medical Teacher* 2017, 39, S50-S54.

48. Scott, K. R.; King, A. M.; Estes, M. K.; Conlon, L. W.; Jones, J. S.; Phillips, A. W. Evaluation of an Intervention to Improve Quality of Single-best Answer Multiple-choice Questions. *Western Journal of Emergency Medicine* 2019, 20, 11-14.

49. Pickering, G. Against Multiple-Choice Questions - Controversy. *Medical Teacher* 1979, 1, 84-86.

50. Treagust, D. F. Development and Use of Diagnostic-Tests to Evaluate Students Misconceptions in Science. *International Journal of Science Education* 1988, 10, 159-169.

51. Cloonan, C. A.; Hutchinson, J. S. A Chemistry Concept Reasoning Test. *Chemistry Education Research and Practice* 2011, 12, 205-209.

52. Smith, T. I.; Louis, K. J.; Ricci, B. J.; Bendjilali, N. Quantitatively ranking incorrect responses to multiple-choice questions using item response theory. *Physical Review Physics Education Research* 2020, 16, 1-16.

53. Nickerson, R. S.; Butler, S. F.; Carlin, M. T. Knowledge Assessment: Squeezing Information From Multiple-Choice Testing. *Journal of Experimental Psychology-Applied* 2015, 21, 167-177.

54. Friel, S.; Johnstone, A. H. Scoring Systems Which Allow for Partial Knowledge. *Journal of Chemical Education* 1978, 55, 717-719.

55. Costello, E.; Holland, J.; Kirwan, C. The future of online testing and assessment: question quality in MOOCs. *International Journal of Educational Technology in Higher Education* 2018, 15, 1-14.

56. Velan, G. M.; Jones, P.; McNeil, H. P.; Kumar, R. K. Integrated Online Formative Assessments in the Biomedical Sciences for Medical Students: Benefits for Learning. *Bmc Medical Education* 2008, 8, 52.

57. Nardi, A.; Ranieri, M. Comparing paper-based and electronic multiple-choice examinations with personal devices: Impact on students' performance, self-efficacy and satisfaction. *British Journal of Educational Technology* 2019, 50, 1495-1506.

58. Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice* 1998, 5, 7-74.

59. Tetteh, G. A.; Sarpong, F. A.-A. Influence of Type of Assessment and Stress on the Learning Outcome. *Journal of International Education in Business* 2015, 8, 125-144.

60. Martinez, M. E. Cognition and the Question of Test Item Format. *Educational Psychologist* 1999, 34, 207-218.

61. Melovitz Vasan, C. A.; DeFouw, D. O.; Holland, B. K.; Vasan, N. S. Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. *Anatomical Sciences Education* 2018, 11, 254-261.

62. Barnett-Foster, D.; Nagy, P. Undergraduate student response strategies to test questions of varying format. *Higher Education* 1996, 32, 177-198.

63. Cobb, K. A.; Brown, G.; Jaarsma, D. A.; Hammond, R. A. The educational impact of assessment: a comparison of DOPS and MCQs. *Medical Teacher* 2013, 35, e1598-607.

64. Harden, R. M. Assessment of Practical Skills: The Objective Structured Practical Examination (OSPE). *Studies in Higher Education* 1980, 5, 187-196.

65. Mahmood, H. Correlation of MCQ and SEQ scores in written undergraduate ophthalmology assessment. *Journal of the College of Physicians and Surgeons Pakistan* 2015, 25, 185.

66. Hudson, R. D.; Treagust, D. F. Which Form of Assessment Provides the Best Information about Student Performance in Chemistry Examinations? *Research in Science & Technological Education* 2013, 31, 49-65.

67. Rodriguez, M. C. Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement* 2003, 40, 163-184.

68. Bacon, D. R. Assessing Learning Outcomes: A Comparison of Multiple-Choice and Short-Answer Questions in a Marketing Context. *Journal of Marketing Education* 2003, 25, 31-36.

69. Azevedo, J. M.; Oliveira, E. P.; Beites, P. D. Using Learning Analytics to evaluate the quality of multiple-choice questions A perspective with Classical Test Theory and Item Response Theory. *International Journal of Information and Learning Technology* 2019, 36, 322-341.

70. Bloom, B. S. *Taxonomy of educational objectives : the classification of educational goals*. Longman Group: London, 1956.

71. Krathwohl, D. R. A revision of Bloom's taxonomy: An overview. *Theory into Practice* 2002, 41, 212-218.

72. Biggs, J. B. *Evaluating the quality of learning : the SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press: New York, London, 1982.

73. Huang, V. An Australian study comparing the use of multiple-choice questionnaires with assignments as interim, summative law school assessment. *Assessment & Evaluation in Higher Education* 2016, 42, 580-595.

74. Liou, P. Y.; Bulut, O. The Effects of Item Format and Cognitive Domain on Students' Science Performance in TIMSS 2011. *Research in Science Education* 2020, 50, 99-121.

75. Considine, J.; Botti, M.; Thomas, S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian* 2005, 12, 19-24.

76. Danili, E.; Reid, N. Assessment Formats: Do They Make a Difference? *Chemistry Education Research and Practice* 2005, 6, 204-212.

77. Downing, S. M. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education* 2005, 10, 133-43.

78. Drost, E. A. Validity and Reliability in Social Science Research. *Education Research and perspectives* 2011, 38, 105.

79. Cronbach, L. J.; Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin* 1955, 52, 281.

80. McKenna, P. Multiple choice questions: answering correctly and knowing the answer. *Interactive Technology and Smart Education* 2019, 16, 59-73.

81. Pham, H.; Trigg, M.; Wu, S. P.; O'Connell, A.; Harry, C.; Barnard, J.; Devitt, P. Choosing Medical Assessments: Does the Multiple-choice Question make the Grade? *Education for Health* 2018, 31, 65-71.

82. Simkin, M. G.; Kuechler, W. L. Multiple-Choice Tests and Student Understanding: What Is the Connection? *Decision Sciences-Journal of Innovative Education* 2005, 3, 73-97.

83. Lederman, N. G.; Abell, S. L. *Handbook of Research on Science Education*. Routledge: 2014.

84. Cole, J. S.; Bergin, D. A.; Whittaker, T. A. Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology* 2008, 33, 609-624.

85. Nicol, D. J.; Macfarlane-Dick, D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education* 2006, 31, 199-218.

86. Wiggins, G. Seven keys to effective feedback. *Feedback* 2012, 70, 10-16.

87. Washburn, S.; Herman, J.; Stewart, R. Evaluation of performance and perceptions of electronic vs. paper multiple-choice exams. *Advances in Physiology Education* 2017, 41, 548-555.

88. Reid, W. A.; Duvall, E.; Evans, P. Relationship between assessment results and approaches to learning and studying in Year Two medical students. *Medical Education* 2007, 41, 754-762.

89. Scouller, K. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education* 1998, 35, 453-472.

90. Marton, F.; Säljö, R. On Qualitative Differences in Learning: I - Outcome and Process. *British Journal of Educational Psychology* 1976, 46, 4-11.

91. Gay, L. R. The Comparative Effects of Multiple-Choice Versus Short-Answer Tests on Retention. *Journal of Educational Measurement* 1980, 17, 45-50.

92. Glover, J. A. The" testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology* 1989, 81, 392.

93. Middlebrooks, C. D.; Murayama, K.; Castel, A. D. Test expectancy and memory for important information. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 2017, 43, 972-985.

94. Roediger, H. L.; Karpicke, J. D. Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science* 2006, 17, 249-255.

95. Alauddin, M.; Khan, A. Does Performance in Progressive Assessment Influence the Outcome in Final Examination? An Australian Experience. *Educational Assessment, Evaluation and Accountability* 2010, 22, 293-305.

96. Kember, D.; Gow, L. A model of student approaches to learning encompassing ways to influence and change approaches. *Instructional Science* 1989, 18, 263-288.

97. Teichert, M. A.; Schroeder, M. J.; Lin, S.; Dillner, D. K.; Komperda, R.; Bunce, D. M. Problem-Solving Behaviors of Different Achievement Groups on Multiple-Choice Questions in General Chemistry. *Journal of Chemical Education* 2020, 97, 3-15.

98. Dolmans, D. H. J. M.; Loyens, S. M. M.; Marcq, H.; Gijbels, D. Deep and surface learning in problem-based learning: a review of the literature. *Advances in Health Sciences Education* 2016, 21, 1087-1112.

99. Brown, S.; White, S.; Wakeling, L.; Naiker, M. Approaches and study skills inventory for students (ASSIST) in an introductory course in chemistry. *Journal of University Teaching & Learning Practice* 2015, 12.

100. Gurpinar, E.; Kulac, E.; Tetik, C.; Akdogan, I.; Mamakli, S. Do learning approaches of medical students affect their satisfaction with problem-based learning? *Advances in Physiology Education* 2013, 37, 85-88.

101. Donnison, S.; Penn-Edwards, S. Focusing on First Year Assessment: Surface or Deep Approaches to Learning? *The International Journal of the First Year in Higher Education* 2012, 3, 9-20.

102. Gijbels, D.; Coertjens, L.; Vanthournout, G.; Struyf, E.; Van Petegem, P. Changing students' approaches to learning: a two-year study within a university teacher training course. *Educational Studies* 2009, 35, 503-513.

103. Segers, M.; Nijhuis, J.; Gijselaers, W. Redesigning a learning and assessment environment: The influence on students' perceptions of assessment demands and their learning strategies. *Studies in Educational Evaluation* 2006, 32, 223-242.

104. Vermunt, J. D. Metacognitive, cognitive and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education* 1996, 31, 25-50.

105. Volet, S. E.; Chalmers, D. Investigation of Qualitative Differences in University Students' Learning Goals, Based on an Unfolding Model of Stage Development. *British Journal of Educational Psychology* 1992, 62, 17-34.

106. Lundeberg, M. A.; Fox, P. W. Do Laboratory Findings on Test Expectancy Generalize to Classroom Outcomes? *Review of Educational Research* 1991, 61, 94-106.

107. Biggs, J. B. Assessing student approaches to learning. *Australian Psychologist* 1988, 23, 197-206.

108. Doğan, C. D.; Atmaca, S.; Yolcu, F. A. The correlation between learning approaches and assessment preferences of eighth-grade students. *İlköğretim Online* 2012, 11.

109. van de Watering, G.; Gijbels, D.; Dochy, F.; van der Rijt, J. Students' assessment preferences, perceptions of assessment and their relationships to study results. *Higher Education* 2008, 56, 645.

110. Kaipa, R. M. Multiple choice questions and essay questions in curriculum. *Journal of Applied Research in Higher Education* 2020, 1-17.

111. Birenbaum, M.; Feldman, R. A. Relationships Between Learning Patterns and Attitudes Towards Two Assessment Formats. *Educational Research* 1998, 40, 90-98.

112. Marshall, G.; Jones, N. A pilot study into the anxiety induced by various assessment methods. *Radiography* 2003, 9, 185-191.

113. Putwain, D. W. Test anxiety and GCSE performance: the effect of gender and socio-economic background. *Educational Psychology in Practice* 2008, 24, 319-334.

114. Ross, M. E.; Salisbury-Glennon, J. D.; Guarino, A.; Reed, C. J.; Marshall, M. Situated self-regulation: Modeling the interrelationships among instruction, assessment, learning strategies and academic performance. *Educational Research and Evaluation* 2003, 9, 189-209.

115. Malau-Aduli, B. S.; Preston, R.; Adu, M.; Alele, F.; Gratani, M.; Drovandi, A.; Heslop, I. Pharmacy students' perceptions of assessment and its impact on learning. *Currents in Pharmacy Teaching and Learning* 2019, 11, 571-579.

116. Salamonson, Y.; Andrew, S.; Everett, B. Academic engagement and disengagement as predictors of performance in pathophysiology among nursing students. *Contemporary Nurse* 2009, 32, 123-132.

117. Lund Dean, K.; Jolly, J. P. Student identity, disengagement, and learning. *Academy of Management Learning & Education* 2012, 11, 228-243.

118. Martin, A. J.; Anderson, J.; Bobis, J.; Way, J.; Vellar, R. Switching on and switching off in mathematics: An ecological study of future intent and disengagement among middle school students. *Journal of Educational Psychology* 2012, 104, 1.

119. Lee, J.-S. The Relationship Between Student Engagement and Academic Performance: Is It a Myth or Reality? *The Journal of Educational Research* 2014, 107, 177-185.

120. Harbour, K. E.; Evanovich, L. L.; Sweigart, C. A.; Hughes, L. E. A Brief Review of Effective Teaching Practices That Maximize Student Engagement. *Preventing School Failure: Alternative Education for Children and Youth* 2015, 59, 5-13.

121. Coates, H. The value of student engagement for higher education quality assurance. *Quality in Higher Education* 2005, 11, 25-36.

122. Krause, K. L.; Coates, H. Students' engagement in first-year university. *Assessment & Evaluation in Higher Education* 2008, 33, 493-505.

123. Finn, J. D.; Rock, D. A. Academic success among students at risk for school failure. *Journal of Applied Psychology* 1997, 82, 221-234.

124. Clifton, R. A.; Baldwin, W. G.; Wei, Y. Course Structure, Engagement, and the Achievement of Students in First-Year Chemistry. *Chemistry Education Research and Practice* 2012, 13, 47-52.

125. D'Mello, S.; Lehman, B.; Pekrun, R.; Graesser, A. Confusion can be beneficial for learning. *Learning and Instruction* 2014, 29, 153-170.

126. Papenberg, M.; Musch, J. Of Small Beauties and Large Beasts: The Quality of Distractors on Multiple-Choice Tests Is More Important Than Their Quantity. *Applied Measurement in Education* 2017, 30, 273-286.

127. Rogers, W. T.; Yang, P. Test-Wiseness: Its Nature and Application. *European Journal of Psychological Assessment* 1996, 12, 247-259.

128. Kim, Y. Partial Identification of Answer Reviewing Effects in Multiple-Choice Exams. *Journal of Educational Measurement* 2019, 1-16.

129. Sarnacki, R. E. An Examination of Test-Wiseness In the Cognitive Test Domain. *Review of Educational Research* 1979, 49, 252-279.

130. Ames, C. Classroom: Goals, Structures, and Student Motivation. *Journal of Educational Psychology* 1992, 84, 261-271.

131. Curtis, D. A.; Lind, S. L.; Boscardin, C. K.; Dellinges, M. Does student confidence on multiple-choice question assessments provide useful information? *Medical Education* 2013, 47, 578-84.

132. Liu, O. L.; Lee, H.-S.; Linn, M. C. Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching* 2011, 48, 1079-1107.

133. Chiavaroli, N. Negatively-Worded Multiple Choice Questions: An Avoidable Threat to Validity. *Practical Assessment, Research & Evaluation* 2017, 22, 1-14.

134. Haladyna, T. M.; Downing, S. M. Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice* 2004, 23, 17-27.

135. Tai, R. H.; Sadler, P. M.; Loehr, J. F. Factors Influencing Success in Introductory College Chemistry. *Journal of Research in Science Teaching* 2005, 42, 987-1012.

136. Sirin, S. R. Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research* 2005, 75, 417-453.

137. McConney, A.; Perry, L. B. Science and Mathematics Achievement in Australia: The Role of School Socioeconomic Composition in Educational Equity and Effectiveness. *International Journal of Science and Mathematics Education* 2010, 8, 429-452.

138. Cunningham, W. G.; Sanzo, T. D. Is High-Stakes Testing Harming Lower Socioeconomic Status Schools? *NASSP Bulletin* 2002, 86, 62-75.

139. Korpershoek, H.; Kuyper, H.; van der Werf, G.; Bosker, R. Who Succeeds in Advanced Mathematics and Science Courses? *British Educational Research Journal* 2011, 37, 357-380.

140. Wilson, T. M.; MacGillivray, H. L. Counting on the Basics: Mathematical Skills Among Tertiary Entrants. *International Journal of Mathematical Education in Science and Technology* 2007, 38, 19-41.

141. Pintrich, P. R. The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research* 1999, 31, 459-470.

142. Schunk, D. H. Self-Efficacy and Academic Motivation. *Educational Psychologist* 1991, 26, 207-231.

143. R. Pintrich, P.; Schrauben, B. Students' motivational beliefs and their cognitive engagement in classroom academic tasks. In *Student perceptions in the classroom*, Lawrence: 1992; pp 149-183.

144. Hailikari, T. K.; Nevgi, A. How to Diagnose At-Risk Students in Chemistry: The Case of Prior Knowledge Assessment. *International Journal of Science Education* 2010, 32, 2079-2095.

145. Brookover, W. B.; Thomas, S.; Paterson, A. Self-Concept of Ability and School Achievement. *Sociology of Education* 1964, 37, 271-278.

146. Boli, J.; Allen, M. L.; Payne, A. High-Ability Women and Men in Undergraduate Mathematics and Chemistry Courses. *American Educational Research Journal* 1985, 22, 605-626.

147. Clarricoates, K. 'Dinosaurs in the classroom'—A re-examination of some aspects of the 'hiddenrs curriculum in primary schools. *Women's Studies International Quarterly* 1978, 1, 353-364.

148. Miller, D. I.; Halpern, D. F. The New Science of Cognitive Sex Differences. *Trends in Cognitive Sciences* 2014, 18, 37-45.

149. Halpern, D. F. Sex Differences in Intelligence - Implications for Education. *American Psychologist* 1997, 52, 1091-1102.

150. Souchal, C.; Toczek, M.-C.; Darnon, C.; Smeding, A.; Butera, F.; Martinot, D. Assessing does not mean threatening: The purpose of assessment as a key determinant of girls' and boys' performance in a science class. *British Journal of Educational Psychology* 2014, 84, 125-136.

151. Kacprzyk, J.; Parsons, M.; Maguire, P. B.; Stewart, G. S. Examining gender effects in different types of undergraduate science assessment. *Irish Educational Studies* 2019, 38, 467-480.

152. Siegfried, C.; Wuttke, E. Are Multiple-Choice Items Unfair? And if So, for Whom? *Citizenship, Social and Economics Education* 2019, 18, 198-217.

153. Halpern, D. F. It's Complicated-In Fact, It's Complex: Explaining the Gender Gap in Academic Achievement in Science and Mathematics. *Psychological Science in the Public Interest* 2014, 15, 72-74.

154. Halpern, D. F.; Benbow, C. P.; Geary, D. C.; Gur, R. C.; Hyde, J. S.; Gernsbacher, M. A. The Science of Sex Differences in Science and Mathematics. *Psychological Science in the Public Interest* 2007, 8, 1-51.

155. Wilson, K.; Low, D.; Verdon, M.; Verdon, A. Differences in Gender Performance on Competitive Physics Selection Tests. *Physical Review Physics Education Research* 2016, 12, 1-16.

156. Low, D.; Malik, U.; Wilson, K. Up and down, but also forward: addressing gender differences in the understanding of projectile motion. *Teaching Science* 2018, 64, 14-20.

157. Cousins, A.; Mills, M. Gender and High School Chemistry: Student Perceptions on Achievement in a Selective Setting. *Cambridge Journal of Education* 2015, 45, 187-204.

158. Barone, C. Some Things Never Change:Gender Segregation in Higher Education across Eight Nations and Three Decades. *Sociology of Education* 2011, 84, 157-176.

159. Alon, S.; Gelbgiser, D. The female advantage in college academic achievements and horizontal sex segregation. *Social Science Research* 2011, 40, 107-119.

160. Mann, A.; Diprete, T. A. Trends in gender segregation in the choice of science and engineering majors. *Social Science Research* 2013, 42, 1519-1541.

161. Ceci, S. J.; Ginther, D. K.; Kahn, S.; Williams, W. M. Women in Academic Science: A Changing Landscape. *Psychological Science in the Public Interest* 2014, 15, 75-141.

162. Turner, R. C.; Lindsay, H. A. Gender Differences in Cognitive and Noncognitive Factors Related to Achievement in Organic Chemistry. *Journal of Chemical Education* 2003, 80, 563-568.

163. Hyde, J. S.; Fennema, E.; Lamon, S. J. Gender Differences in Mathematics Performance: A Meta-Analysis. *Psychological Bulletin* 1990, 107, 139-155.

164. Samuelsson, M.; Samuelsson, J. Gender differences in boys' and girls' perception of teaching and learning mathematics. *Open Review of Educational Research* 2016, 3, 18-34.

165. Maries, A.; Karim, N. I.; Singh, C. Is agreeing with a gender stereotype correlated with the performance of female students in introductory physics? *Physical Review Physics Education Research* 2018, 14, 1-10.

166. Stricker, L. J.; Rock, D. A.; Burton, N. W. Sex differences in predictions of college grades from scholastic aptitude test scores. *Journal of Educational Psychology* 1993, 85, 710.

167. Spencer, S. J.; Steele, C. M.; Quinn, D. M. Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology* 1999, 35, 4-28.

168. Steele, C. M. A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist* 1997, 52, 613.

169. Stieff, M. Sex Differences in the Mental Rotation of Chemistry Representations. *Journal of Chemical Education* 2013, 90, 165-170.

170. Walding, R.; Fogliani, C.; Over, R.; Bain, J. D. Gender Differences in Response to Questions on the Australian National Chemistry Quiz. *Journal of Research in Science Teaching* 1994, 31, 833-846.

171. Duckworth, A. L.; Seligman, M. E. P. Self-Discipline Gives Girls the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores. *Journal of Educational Psychology* 2006, 98, 198-208.

172. Hamilton, L. S. Gender Differences on High School Science Achievement Tests: Do Format and Content Matter? *Educational Evaluation and Policy Analysis* 1998, 20, 179-195.

173. Hudson, R. D. Is There a Relationship Between Chemistry Performance and Question Type, Question Content and Gender? *Science Education International* 2012, 23, 56-83.

174. Hedgeland, H.; Dawkins, H.; Jordan, S. Investigating Male Bias in Multiple Choice Questions: Contrasting Formative and Summative Settings. *European Journal of Physics* 2018, 39, 1-7.

175. Beard, J.; Fogliani, C.; Owens, C.; Wilson, A. Is Achievement in Australian Chemistry Gender Based? *Research in Science Education* 1993, 23, 10-14.

176. Cox, P. J.; Leder, G. C.; Forgasz, H. J. Victorian Certificate of Education: Mathematics, Science and Gender. *Australian Journal of Education* 2004, 48, 27-46.

177. Reardon, S. F.; Kalogrides, D.; Fahle, E. M.; Podolsky, A.; Zárate, R. C. The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades. *Educational Researcher* 2018, 47, 284-294.

178. Buccheri, G.; Gurber, N. A.; Bruhwiler, C. The Impact of Gender on Interest in Science Topics and the Choice of Scientific and Technical Vocations. *International Journal of Science Education* 2011, 33, 159-178.

179. Moss-Racusin, C. A.; Dovidio, J. F.; Brescoll, V. L.; Graham, M. J.; Handelsman, J. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* 2012, 109, 16474-16479.

180. Kelly, F. J. The Kansas Silent Reading Tests. *Journal of Educational Psychology* 1916, 7, 63-80.

181. Afifi, M.; Hussain, K. F. The achievement of higher flexibility in multiple-choice-based tests using image classification techniques. *International Journal on Document Analysis and Recognition* 2019, 22, 127-142.

182. Anderson, J. The MCQ Controversy - A Review. *Medical Teacher* 1981, 3, 150-156.

183. Atalmis, E. H.; Kingston, N. M. The Impact of Homogeneity of Answer Choices on Item Difficulty and Discrimination. *SAGE Open* 2018, 8, 1-9.

184. BarnettFoster, D.; Nagy, P. Undergraduate student response strategies to test questions of varying format. *Higher Education* 1996, 32, 177-198.

185. Kiat, J. E.; Ong, A. R.; Ganesan, A. The influence of distractor strength and response order on MCQ responding. *Educational Psychology* 2017, 38, 368-380.

186. Carnegie, J. A. Does Correct Answer Distribution Influence Student Choices When Writing Multiple Choice Examinations? *Canadian Journal for the Scholarship of Teaching and Learning* 2017, 8, 1-16.

187. Tellinghuisen, J.; Sulikowski, M. M. Does the Answer Order Matter on Multiple-Choice Exams? *Journal of Chemical Education* 2008, 85, 572.

188. Fozzard, N.; Pearson, A.; du Toit, E.; Naug, H.; Wen, W.; Peak, I. R. Analysis of MCQ and distractor use in a large first year Health Faculty Foundation Program: assessing the effects of changing from five to four options. *BMC Medical Education* 2018, 18, 1-10.

189. Schneid, S. D.; Armour, C.; Park, Y. S.; Yudkowsky, R.; Bordage, G. Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. *Medical Education* 2014, 48, 1020-7.

190. Raymond, M. R.; Stevens, C.; Bucak, S. D. The optimal number of options for multiple-choice questions on high-stakes tests: application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education* 2019, 24, 141-150.

191. Vegada, B.; Shukla, A.; Khilnani, A.; Charan, J.; Desai, C. Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian Journal of Pharmacology* 2016, 48, 571-575.

192. Edwards, B. D.; Arthur Jr, W.; Bruce, L. L. The Three-option Format for Knowledge and Ability Multiple-choice Tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment* 2012, 20, 65-81.

193. Dressel, P. L.; Schmid, J. Some Modifications of the Multiple-Choice Item. *Educational and Psychological Measurement* 1953, 13, 574-595.

194. Veloski, J. J.; Rabinowitz, H. K.; Robeson, M. R. A Solution to the Cueing Effects of Multiple-Choice Questions - The UN-Q Format. *Medical Education* 1993, 27, 371-375.

195. Draper, S. W. Catalytic Assessment: Understanding how MCQs and EVS can Foster Deep Learning. *British Journal of Educational Technology* 2009, 40, 285-293.

196. Tweed, M.; Wilkinson, T. A randomized controlled trial comparing instructions regarding unsafe response options in a MCQ examination. *Medical Teacher* 2009, 31, 51-54.

197. Davies, G. R.; Proops, H.; Carolan, C. M. The Development and Use of a Multiple-Choice Question (MCQ) Assessment to Foster Deeper Learning: An Exploratory Web-Based Qualitative Investigation. *Journal of Teaching and Learning* 2020, 14, 1-12.

198. Hsia, Y. T.; Jong, B. S.; Lin, T. W.; Liao, J. Y. Designating "hot" items in multiple-choice questions-A strategy for reviewing course materials. *Journal of Computer Assisted Learning* 2019, 35, 188-196.

199. Collignon, S. E.; Chacko, J.; Martin, M. W. An Alternative Multiple-Choice Question Format to Guide Feedback Using Student Self-Assessment of Knowledge. *Decision Sciences-Journal of Innovative Education* 2020, 1-25.

200. Couch, B. A.; Brassil, C. E.; Hubbard, J. K. Multiple–True–False Questions Reveal the Limits of the Multiple–Choice Format for Detecting Students with Incomplete Understandings. *BioScience* 2018, 68, 455-463.

201. Scharf, E. M.; Baldwin, L. P. Assessing multiple choice question (MCQ) tests - a mathematical perspective. *Active Learning in Higher Education* 2007, 8, 31-47.

202. Romm, A. T.; Schoer, V.; Kika, J. C. A test taker's gamble: The effect of average grade to date on guessing behaviour in a multiple choice test with a negative marking rule. *South African Journal of Economic and Management Sciences* 2019, 22, 12.

203. Frary, R. B. Formula Scoring of Multiple-Choice Tests (Correction for Guessing). *Educational Measurement: Issues and Practice* 1988, 7, 33-38.

204. Aray, H.; Pedauga, L. The Value of Choice: An Experiment Using Multiple-Choice Tests. *Educational Measurement-Issues and Practice* 2019, 38, 20-32.

205. Jalloh, C.; Collins, B.; Lafleur, D.; Reimer, J.; Morrow, A. Mapping session learning objectives to exam questions: How to do it and how to apply the results. *Medical Teacher* 2019, 1-7.

206. M. Tiemeier, A.; Stacy, Z.; M. Burke, J. Using Multiple Choice Questions Written at Various Bloom's Taxonomy Levels to Evaluate Student Performance across a Therapeutics Sequence. *Innovations in Pharmacy* 2011, 2.

207. Hartman, J. R.; Lin, S. Analysis of Student Performance on Multiple-Choice Questions in General Chemistry. *Journal of Chemical Education* 2011, 88, 1223-1230.

208. Melser, M. C.; Steiner-Hofbauer, V.; Lilaj, B.; Agis, H.; Knaus, A.; Holzinger, A. Knowledge, application and how about competence? Qualitative assessment of multiple-choice questions for dental students. *Medical Education Online* 2020, 25, 1-8.

209. Nicol, D. E-Assessment by Design: Using Multiple-Choice Tests to Good Effect. *Journal of Further and Higher Education* 2007, 31, 53-64.

210. Buckles, S.; Siegfried, J. J. Using Multiple-Choice Questions to Evaluate In-Depth Learning of Economics. *Journal of Economic Education* 2006, 37, 48-57.

211. Klopfer, L. E. A Structure for the Affective Domain in Relation to Science Education. *Science education : Australian Practices and Perspectives* 1976, 60, 299.

212. Kauertz, A.; Fischer, H. E. Assessing Students' Level of Knowledge and Analysing the Reasons for Learning Difficulties in Physics by Rasch Analysis. *Applications of Rasch measurement in Science Education* 2006, 212-246.

213. Claesgens, J.; Scalise, K.; Wilson, M.; Stacy, A. Mapping student understanding in chemistry: The Perspectives of Chemists. *Science Education* 2009, 93, 56-85.

214. Smith, K. C.; Nakhleh, M. B.; Bretz, S. L. An Expanded Framework for Analyzing General Chemistry Exams. *Chemistry Education Research and Practice* 2010, 11, 147-153.

215. Lee, H.-S.; Liu, O. L.; Linn, M. C. Validating Measurement of Knowledge Integration in Science Using Multiple-Choice and Explanation Items. *Applied Measurement in Education* 2011, 24, 115-136.

216. Chandler, P.; Sweller, J. Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction* 1991, 8, 293-332.

217. Pachai, M. V.; DiBattista, D.; Kim, J. A. A Systematic Assessment of 'None of the Above' on Multiple Choice Tests in a First Year Psychology Classroom. *Canadian Journal for the Scholarship of Teaching and Learning* 2015, 1-17.

218. Azer, S. A. Assessment in a Problem-Based Learning Course - Twelve Tips for Constructing Multiple Choice Questions that Test Students' Cognitive Skills. *Biochemistry and Molecular Biology Education* 2003, 31, 428-434.

219. Towns, M. H. Guide To Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments. *Journal of Chemical Education* 2014, 91, 1426-1431.

220. Bernhofer, E. I.; Burchill, C. N. Developing and Evaluating Multiple-Choice Tests for Trustworthiness. *Journal for Nurses in Professional Development* 2019, 35, 204-209.

221. Cox, C. W. Best Practice Tips for the Assessment of Learning of Undergraduate Nursing Students via Multiple-Choice Questions. *Nursing Education Perspectives* 2019, 40, 228-230.

222. Breakall, J.; Randles, C.; Tasker, R. Development and Use of a Multiple-Choice Item Writing Flaws Evaluation Instrument in the Context of General Chemistry. *Chemistry Education Research and Practice* 2019, 20, 369-382.

223. Riccardi, D.; Lightfoot, J.; Lam, M.; Lyon, K.; Roberson, N. D.; Lolliot, S. Investigating the effects of reducing linguistic complexity on EAL student comprehension in first-year undergraduate assessments. *Journal of English for Academic Purposes* 2020, 43, 1-11.

224. Ali, S. H.; Carr, P. A.; Ruit, K. G. Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter. *Journal of the Scholarship of Teaching and Learning* 2016, 16, 1-14.

225. Mulford, D. R.; Robinson, W. R. An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *Journal of Chemical Education* 2002, 79, 739-744.

226. Shin, J.; Bulut, O.; Gierl, M. J. The Effect of the Most-Attractive-Distractor Location on Multiple-Choice Item Difficulty. *Journal of Experimental Education* 2019, 1-17.

227. Holzknecht, F.; McCray, G.; Eberharter, K.; Kremmel, B.; Zehentner, M.; Spiby, R.; Dunlea, J. The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing* 2020, 1-21.

228. Wang, L. *Does Rearranging Multiple-Choice Item Response Options Affect Item and Test Performance?* ; Educational Testing Service: 2019.

229. Kurdi, G.; Leo, J.; Parsia, B.; Sattler, U.; Al-Emari, S. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education* 2020, 30, 121-204.

230. Rao, C. H. D.; Saha, S. K. Automatic Multiple Choice Question Generation From Text: A Survey. *Ieee Transactions on Learning Technologies* 2020, 13, 14-25.

231. Sad, S. N. Does difficulty-based item order matter in multiple-choice exams? (Empirical evidence from university students). *Studies in Educational Evaluation* 2020, 64, 1-8.

232. Ding, L.; Beichner, R. Approaches to Data Analysis of Multiple-Choice Questions. *Physical Review Special Topics - Physics Education Research* 2009, 5.

233. Stouffer, S. A.; Guttman, L.; Suchman, E. A.; Lazarsfeld, P. F.; Star, S. A.; Clausen, J. A. *Measurement and Prediction*. Princeton University Press: 1950.

234. Wright, B. D. Estimating Rasch Measures for Extreme Scores. *Rasch Measurement Transactions* 1998, 12:2, 162-633.

235. Belouafa, S.; Habti, F.; Benhar, S.; Belafkih, B.; Tayane, S.; Hamdouch, S.; Bennamara, A.; Abourriche, A. Statistical tools and approaches to validate analytical methods: methodology and practical examples. *International Journal of Metrology and Quality Engineering* 2017, 8, 1-10.

236. Strasser, N. Avoiding Statistical Mistakes. *Journal of College Teaching & Learning* 2007, 4, 51-58.

237. De Veaux, R. D. *Intro stats*. 4th ed. ed.; Pearson: Boston, Mass., 2014.

238. Fischer, H. *A history of the central limit theorem from classical to modern probability theory*. Springer: New York, 2011.

239. Springate, S. D. The effect of sample size and bias on the reliability of estimates of error: a comparative study of Dahlberg's formula. *European Journal of Orthodontics* 2011, 34, 158-163.

240. Biau, D. J.; Kernéis, S.; Porcher, R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clinical Orthopaedics and Related Research* 2008, 466, 2282-2288.

241. Anthoine, E.; Moret, L.; Regnault, A.; Sébille, V.; Hardouin, J.-B. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health and quality of life outcomes* 2014, 12, 176-176.

242. Altman, D. G.; Bland, J. M. Statistics notes: The normal distribution. *BMJ* 1995, 310, 298.

243. Begg, M. D. Sampling Distributions. In *Encyclopedia of Biostatistics*, 2005; pp 4752-4754.

244. Neyman, J.; Pearson, E. S. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part II. *Biometrika* 1928, 20A, 263-294.

245. Neyman, J.; Pearson, E. S. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika* 1928, 20A, 175-240.

246. Wilk, M. B.; Gnanadesikan, R. Probability Plotting Methods for the Analysis of Data. *Biometrika* 1968, 55, 1-17.

247. Student. The Probable Error of a Mean. *Biometrika* 1908, 6, 1-25.

248. Brown, L. D.; Cai, T. T.; DasGupta, A. Interval Estimation for a Binomial Proportion. *Statistical Science* 2001, 16, 101-133.

249. Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association* 1927, 22, 209-212.

250. Welch, B. L. The generalisation of student's problems when several different population variances are involved. *Biometrika* 1947, 34, 28-35.

251. John, B. D. V. W. Significance of the Difference between Two Means when the Population Variances may be Unequal. *Nature* 1960, 187, 438.

252. Kim, T. K. Understanding one-way ANOVA using conceptual figures. *Korean Journal of Anesthesiology* 2017, 70, 22-26.

253. Vickers, A. J. Analysis of variance is easily misapplied in the analysis of randomized trials: a critique and discussion of alternative statistical approaches. *Psychosomatic Medicine* 2005, 67, 652-5.

254. Delucchi, K. L. The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin* 1983, 94, 166-176.

255. Wilson, E. B.; Hilferty, M. M. The Distribution of Chi-square. *Proceedings of the National Academy of Sciences of the United States of America* 1931, 17, 684-688.

256. McHugh, M. L. The chi-square test of independence. *Biochemia Medica* 2013, 23, 143-149.

257. Cohen, J. Statistical Power Analysis. *Current Directions in Psychological Science* 1992, 1, 98-101.

258. Cohen, J. *Statistical power analysis for the behavioral sciences*. Routledge: 2013.

259. Dunn, O. J. Estimation of the Medians for Dependent Variables. *Annals of Mathematical Statistics* 1959, 30, 192-197.

260. Dunn, O. J. Multiple Comparisons among Means. *Journal of the American Statistical Association* 1961, 56, 52-64.

261. Miller, R. G. *Simultaneous statistical inference*. McGraw-Hill: New York, 1966.

262. Shaffer, J. P. Multiple Hypothesis-Testing. *Annual Review of Psychology* 1995, 46, 561-584.

263. Novick, M. R. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology* 1966, 3, 1-18.

264. Traub, R. E. Classical Test Theory in Historical Perspective. *Educational Measurement: Issues and Practice* 1997, 16, 8-14.

265. Kline, T. J. B. *Psychological testing: A practical approach to design and evaluation*. Sage Publications, Inc: Thousand Oaks, CA, US, 2005; p xii, 356-xii, 356.

266. DeVellis, R. F. Classical Test Theory. *Medical Care* 2006, 44, S50-S59.

267. Crocker, L. M.; Algina, J. *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston: 1986.

268. Sijtsma, K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 2009, 74, 107.

269. Sullivan, G. M.; Feinn, R. Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education* 2012, 4, 279-282.

270. Ding, L.; Chabay, R.; Sherwood, B.; Beichner, R. Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research* 2006, 2, 010105.

271. Young, M.; Cummings, B. A.; St-Onge, C. Ensuring the quality of multiple-choice exams administered to small cohorts: A cautionary tale. *Perspectives on Medical Education* 2017, 6, 21-28.

272. Rasch, G. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche: Oxford, England, 1960; p xiii, 184-xiii, 184.

273. Bond, T. G. F., Christine M. *Applying the Rasch model : fundamental measurement in the human sciences*. 2nd ed. ed.; Lawrence Erlbaum Associates: Mahwah, NJ, 2007.

274. Boone, W. J. Rasch Analysis for Instrument Development: Why, When, and How? *CBE: Life Sciences Education* 2016, 15, 1-7.

275. Ziegler, M.; Hagemann, D. Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment* 2015, 31, 231-237.

276. Wright, B. D.; Linacre, J. M. Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation* 1989, 70, 857-60.

277. Tennant, A.; Pallant, J. Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions* 2006, 20, 1048-1051.

278. Smith, J. E. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement* 2002, 3, 205-231.

279. Merbitz, C.; Morris, J.; Grip, J. C. Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation* 1989, 70, 308-312.

280. Wright, B. D. Thinking with Raw Scores. *Rasch Measurement Transactions* 1993, 7:2, 299-300.

281. Doucette, A.; Wolf, A. W. Questioning the measurement precision of psychotherapy research. *Psychotherapy Research* 2009, 19, 374-389.

282. Masters, G. N. A Rasch model for partial credit scoring. *Psychometrika* 1982, 47, 149-174.

283. Tavakol, M.; Dennick, R. Psychometric Evaluation of a Knowledge Based Examination using Rasch Analysis: An Illustrative Guide: AMEE Guide No. 72. *Medical Teacher* 2013, 35, e838-48.

284. Wright, B.; Panchapakesan, N. A Procedure for Sample-Free Item Analysis. *Educational and Psychological Measurement* 1969, 29, 23-48.

285. Linn, R. L. B. J. A. The Rasch Model, Objective Measurement, Equating, and Robustness. *Applied Psychological Measurement.* 1979, 3, 437.

286. Sondergeld, T. A.; Johnson, C. C. Using Rasch Measurement for the Development and Use of Affective Assessments in Science Education Research. *Science Education* 2014, 98, 581-613.

287. Cunningham, J. D.; Bradley, K. D. In *Applying the Rasch Model to Measure Change in Student Performance over Time*, American Educational Research Association Annual Meeting, Denver, CO, 2010.

288. Yu, C. H. A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling. http:// www.creativewisdom.com, 2016.

289. Smith, R. M. Fit analysis in latent trait measurement models. *Journal of applied measurement* 2000, 1, 199-218.

290. Millman, J.; Bishop, C. H.; Ebel, R. An Analysis of Test-Wiseness. *Educational and Psychological Measurement* 1965, 25, 707-726.

291. Li, A. W. P. *Dictionary of evidence-based Medicine*. Radcliffe Publishing: 1998.

292. Andrich, D. A rating formulation for ordered response categories. *Psychometrika* 1978, 43, 561-573.

293. Rupp, A. A.; Zumbo, B. D. Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement* 2006, 66, 63-84.

294. Hambleton, R. K.; Jones, R. W. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice* 1993, 12, 38-47.

295. Douglass, F. M.; Khavari, K. A.; Farber, P. D. Comparison of Classical and Latent Trait Item Analysis Procedures. *Educational and Psychological Measurement* 1979, 39, 337-352.

296. Boone, W. J.; Scantlebury, K. The Role of Rasch Analysis when Conducting Science Education Research Utilizing Multiple-Choice Tests. *Science Education* 2006, 90, 253-269.

297. Arjoon, J. A.; Xu, X.; Lewis, J. E. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *Journal of Chemical Education* 2013, 90, 536-545.

298. Barbera, J. A Psychometric Analysis of the Chemical Concepts Inventory. *Journal of Chemical Education* 2013, 90, 546-553.

299. Doran, R. L. *Basic measurement and evaluation of science instruction*. National Science Teachers Association: Washington, D.C, 1980.

300. Loevinger, J. The Attenuation Paradox in Test Theory. *Psychological Bulletin* 1954, 51, 493-504.

301. Ghiselli, E. E.; Campbell, J. P.; Zedeck, S. *Measurement Theory for the Behavioral Sciences*. W. H. Freeman: 1981.

302. Cureton, E. E. The upper and lower twenty-seven per cent rule. *Psychometrika* 1957, 22, 293-296.

303. Kline, P. *A Handbook of Test Construction: Introduction to Psychometric Design*. Methuen: 1986.

304. Kuder, G.; Richardson, M. The theory of the estimation of test reliability. *Psychometrika* 1937, 2, 151-160.

305. Cronbach, L. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951, 16, 297-334.

306. Streiner, D. L. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment* 2003, 80, 99-103.

307. Hankins, M. Questionnaire discrimination: (re)-introducing coefficient delta. *BMC Medical Research Methodology* 2007, 7.

308. Linacre, J. M. *Winsteps® Rasch Measurement Computer Program*, Beaverton, Oregon: Winsteps.com, 2018.

309. Linacre, J. M. Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com, 2018.

310. Linacre, J. M. What do Infit and Outfit, Mean-Square and Standardised mean? *Rasch Measurement Transactions* 2002, 16:2, 878.

311. Wright, B. D. Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement* 1977, 14, 97-116.

312. Masters, G. N. Item Discrimination - When More is Worse. *Journal of Educational Measurement* 1988, 25, 15-29.

313. Harman, H. H. *Modern Factor Analysis*. University of Chicago Press: 1960.

314. Chang, C. H. Finding two dimensions in MMPI-2 depression. *Structural Equation Modeling: A Multidisciplinary Journal* 1996, 3, 41-49.

315. Smith, R. M. A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal* 1996, 3, 25-40.

316. Wright, B. D. Comparing Rasch measurement and factor analysis. *Structural Equation Modeling* 1996, 3, 3-24.

317. González-Espada, W. J. Detecting Gender Bias Through Test Item Analysis. *The Physics Teacher* 2009, 47, 175-179.

318. Mantel, N.; Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959, 22, 719-748.

319. Linacre, J. M. W., B. D. Mantel-Haenszel DIF and PROX are Equivalent! *Rasch Measurement Transactions* 1989, 3:2, 52-53.

320. Engelhard Jr, G. An Empirical Comparison of Mantel-Haenszel and Rasch Procedures for Studying Differential Item Functioning on Teacher Certification Tests. *Journal of Research and Development in Education* 1990, 23, 172-79.

321. Schulz, E. M. DIF Detection: Rasch versus Mantel_Haenszel. *Rasch Measurement Transactions* 1990, 4:2, 107.

322. Matthew, E.; Wright, B. D. An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. *Objective Measurement: Theory into Practice* 1992, 3, 65.

323. Linacre, J. M.; Wright, B. D. *Mantel-Haenszel and the Rasch Model*; University of Chicago: 1987.

324. Bechger, T. M.; Maris, G. A Statistical Test for Differential Item Pair Functioning. *Psychometrika* 2015, 80, 317-40.

325. Andrich, D.; Hagquist, C. Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics* 2012, 37, 387-416.

326. Borsboom, D. When Does Measurement Invariance Matter? *Medical Care* 2006, 44, 176-181.

327. Tennant, A. P., J. F. DIF Matters: A Practical Approach to Test if Differential Item Functioning makes a Difference. *Rasch Measurement Transactions* 2007, 20:4, 1082-84.

328. Wang, W. C. Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education* 2004, 72, 221-261.

329. Kopf, J.; Zeileis, A.; Strobl, C. A Framework for Anchor Methods and an Iterative Forward Approach for DIF Detection. *Applied Psychological Measurement* 2014, 39, 83-103.

330. Kopf, J.; Zeileis, A.; Strobl, C. Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches. *Educational and Psychological Measurement* 2015, 75, 22-56.

331. Wright, B. D. Rack and Stack: Time 1 vs. Time 2 or Pre-Test vs. Post-Test. *Rasch Measurement Transactions* 2003, 17, 905-906.

332. Wright, B. D. Time 1 to Time 2 (Pre-Test to Post-Test) Comparison: Racking and Stacking. *Rasch Measurement Transactions* 1996, 10:1, 478.

333. Kim, H.-Y. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics* 2013, 38, 52-54.

334. Fleming, N. D. *Teaching and Learning Styles: VARK strategies*. IGI global: 2011.

Appendix

7.1 MCQ Assessment Exploratory Analysis and Tests of Normality

Table 41: The MCQ Assessment Results from Chemistry IA from 2012 - 2015 Including the Measures of Spread and the Results of The Tests of Normality Undertaken on those Results

							Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean Score	Std. Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Chemistry IA	2012	Lecture Test 1	471	8.760	3.214	0.148	-0.067	0.113	-0.709	0.225	0.086	470	<<0.001	0.979	470	<<0.001
		Lecture Test 2	446	8.500	2.646	0.125	-0.290	0.116	-0.569	0.231	0.104	446	<<0.001	0.971	446	<<0.001
		Exam Test	508	4.610	2.099	0.093	0.211	0.108	-0.364	0.216	0.114	508	<<0.001	0.973	508	<<0.001
		Redeemable Exam	488	16.090	7.395	0.335	-0.420	0.111	-0.583	0.221	0.079	488	<<0.001	0.954	488	<<0.001
	2013	Lecture Test 1	449	9.130	3.227	0.152	-0.156	0.115	-0.793	0.230	0.104	449	<<0.001	0.973	449	<<0.001
		Lecture Test 2	421	8.950	2.661	0.130	-0.451	0.119	-0.494	0.237	0.134	421	<<0.001	0.958	421	<<0.001
		Exam Test	506	4.750	2.061	0.092	0.100	0.109	-0.435	0.217	0.105	506	<<0.001	0.973	506	<<0.001
		Redeemable Exam	490	16.640	7.601	0.343	-0.477	0.110	-0.554	0.220	0.084	490	<<0.001	0.949	490	<<0.001
	2014	Lecture Test 1	474	8.330	2.844	0.131	0.001	0.112	-0.472	0.224	0.072	474	<<0.001	0.984	474	<<0.001
		Lecture Test 2	436	9.090	2.838	0.136	-0.548	0.117	-0.217	233.000	0.124	436	<<0.001	0.960	436	<<0.001
		Exam Test	509	6.000	2.213	0.098	-0.098	0.108	-0.781	0.216	0.101	509	<<0.001	0.966	509	<<0.001
		Redeemable Exam	499	16.050	7.244	0.324	-0.472	0.109	-0.446	0.218	0.078	499	<<0.001	0.954	499	<<0.001
	2015	Lecture Test 1	504	8.700	2.826	0.126	-0.042	0.109	-0.376	0.217	0.086	504	<<0.001	0.984	504	<<0.001
		Lecture Test 2	451	9.510	2.842	0.134	-0.172	0.115	-0.574	0.229	0.084	451	<<0.001	0.976	451	<<0.001
		Exam Test	547	5.850	2.218	0.095	-0.279	0.104	-0.390	0.209	0.110	547	<<0.001	0.970	547	<<0.001
		Redeemable Exam	525	15.980	6.806	0.297	-0.603	0.107	-0.148	0.213	0.094	525	<<0.001	0.944	525	<<0.001

Table 42: The MCQ Assessment Results from Chemistry IB from 2012 - 2015 Including the Measures of Spread and the Results of The Tests of Normality Undertaken on those Results

							Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean Score	Std. Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Chemistry IB	2012	Lecture Test 1	384	8.930	2.931	0.150	0.143	0.125	-0.712	0.248	0.101	384	<<0.001	0.974	384	<<0.001
		Lecture Test 2	364	7.350	3.187	0.167	0.369	0.128	-0.297	0.255	0.106	364	<<0.001	0.974	364	<<0.001
		Exam Test	434	4.210	1.827	0.088	0.100	0.117	-0.363	0.234	0.124	434	<<0.001	0.971	434	<<0.001
		Redeemable Exam	421	16.240	6.728	0.328	-0.417	0.119	-0.188	0.237	0.072	421	<<0.001	0.967	421	<<0.001
	2013	Lecture Test 1	378	9.280	2.945	0.151	-0.084	0.125	-0.713	0.250	0.096	378	<<0.001	0.977	378	<<0.001
		Lecture Test 2	348	7.850	3.047	0.163	0.106	0.131	-0.720	0.261	0.113	348	<<0.001	0.977	348	<<0.001
		Exam Test	450	5.230	2.041	0.096	-0.115	0.115	-0.496	0.230	0.103	450	<<0.001	0.973	450	<<0.001
		Redeemable Exam	434	16.910	7.352	0.353	-0.494	0.117	-0.425	0.234	0.090	434	<<0.001	0.948	434	<<0.001
	2014	Lecture Test 1	423	9.330	3.229	0.157	-0.103	0.119	-0.807	0.237	0.094	423	<<0.001	0.973	423	<<0.001
		Lecture Test 2	395	7.890	3.137	0.158	0.273	0.123	-0.684	0.245	0.119	395	<<0.001	0.972	395	<<0.001
		Exam Test	486	5.230	2.113	0.096	0.038	0.111	-0.424	0.221	0.103	486	<<0.001	0.975	486	<<0.001
		Redeemable Exam	456	16.890	6.818	0.319	-0.523	0.114	-0.143	0.228	0.077	456	<<0.001	0.957	456	<<0.001
	2015	Lecture Test 1	429	8.990	3.049	0.147	0.109	0.118	-0.766	0.235	0.094	429	<<0.001	0.975	429	<<0.001
		Lecture Test 2	393	8.340	2.881	0.145	0.223	0.123	-0.423	0.246	0.093	393	<<0.001	0.979	393	<<0.001
		Exam Test	487	4.980	2.118	0.096	-0.008	0.111	-0.645	0.221	0.104	487	<<0.001	0.972	487	<<0.001
		Redeemable Exam	472	16.960	7.261	0.334	-0.469	0.112	-0.290	0.224	0.075	472	<<0.001	0.955	472	<<0.001

Table 43: The MCQ Assessment Results from Foundations of Chemistry IA from 2012 - 2015 Including the Measures of Spread and the Results of The Tests of Normality Undertaken on those Results

							Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean Score	Std. Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Foundations of Chemistry IA	**2012**	**Lecture Test 1**	259	11.340	2.407	0.150	-0.542	0.151	0.114	0.302	0.137	259	<<0.001	0.952	259	<<0.001
		Lecture Test 2	267	9.160	2.796	0.171	-0.274	0.149	-0.639	0.297	0.109	267	<<0.001	0.969	267	<<0.001
		Exam Test	306	4.940	2.076	0.119	0.106	0.139	-0.321	0.278	0.100	306	<<0.001	0.975	306	<<0.001
		Redeemable Exam	258	18.490	4.965	0.309	-0.176	0.152	-0.534	0.020	0.074	258	0.002	0.984	258	0.006
	2013	**Lecture Test 1**	309	11.180	2.393	0.136	-0.576	0.139	0.162	0.276	0.133	309	<<0.001	0.955	309	<<0.001
		Lecture Test 2	255	8.600	2.922	0.183	-0.195	0.153	-0.612	0.304	0.096	255	<<0.001	0.979	255	0.001
		Exam Test	365	4.770	2.117	0.111	0.261	0.128	-0.537	0.255	0.121	365	<<0.001	0.967	365	<<0.001
		Redeemable Exam	336	18.890	4.728	0.258	0.081	0.133	-0.563	0.265	0.065	336	0.001	0.986	336	0.003
	2014	**Lecture Test 1**	252	11.200	2.683	0.169	-0.908	0.153	0.769	0.306	0.134	252	<<0.001	0.929	252	<<0.001
		Lecture Test 2	223	9.000	2.664	0.178	-0.152	0.163	-0.608	0.324	0.088	223	<<0.001	0.972	223	<<0.001
		Exam Test	327	4.590	2.127	0.118	0.316	0.135	-0.527	0.269	0.124	327	<<0.001	0.965	327	<<0.001
		Redeemable Exam	301	18.460	5.221	0.301	-0.048	0.140	-0.403	0.280	0.056	301	0.023	0.990	301	0.340
	2015	**Lecture Test 1**	294	11.130	2.690	0.157	-0.667	0.142	0.059	0.283	0.127	294	<<0.001	0.945	294	<<0.001
		Lecture Test 2	236	9.030	2.931	0.191	-0.387	0.158	-0.353	0.316	0.123	236	<<0.001	0.973	236	<<0.001
		Exam Test	367	4.900	2.097	0.109	0.163	0.127	-0.491	0.254	0.107	367	<<0.001	0.972	367	<<0.001
		Redeemable Exam	331	18.380	5.271	0.290	-0.186	0.134	-0.358	0.267	0.067	331	0.001	0.988	331	0.008

Table 44: The MCQ Assessment Results from Foundations of Chemistry IB from 2012 - 2015 Including the Measures of Spread and the Results of The Tests of Normality Undertaken on those Results

							Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean Score	Std. Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Foundations of Chemistry IB	2012	Lecture Test 1	238	10.210	3.017	0.196	-0.434	0.158	-0.460	0.314	0.114	238	<<0.001	0.963	238	<<0.001
		Lecture Test 2	190	7.230	2.933	0.213	-0.101	0.176	-0.606	0.351	0.075	190	0.012	0.978	190	0.007
		Exam Test	268	6.350	1.816	0.111	-0.360	0.149	0.383	0.297	0.133	268	<<0.001	0.961	268	<<0.001
		Redeemable Exam	250	15.830	8.095	0.512	-0.427	0.154	-0.858	0.307	0.112	250	<<0.001	0.933	250	<<0.001
	2013	Lecture Test 1	249	9.470	2.957	0.187	-0.205	0.154	-0.758	0.307	0.100	249	<<0.001	0.969	249	<<0.001
		Lecture Test 2	218	8.830	3.119	0.211	-0.171	0.165	-0.988	0.328	0.119	218	<<0.001	0.961	218	<<0.001
		Exam Test	307	4.010	2.144	0.122	0.390	0.139	-0.294	0.277	0.137	307	<<0.001	0.964	307	<<0.001
		Redeemable Exam	288	16.990	7.247	0.427	-0.689	0.144	-0.269	0.286	0.113	288	<<0.001	0.930	288	<<0.001
	2014	Lecture Test 1	216	9.580	3.160	0.215	-0.410	0.166	-0.514	0.330	0.120	216	<<0.001	0.963	216	<<0.001
		Lecture Test 2	184	9.510	3.046	0.225	-0.165	0.179	-0.512	0.356	0.104	184	<<0.001	0.973	184	0.001
		Exam Test	276	5.070	2.001	0.120	0.344	0.147	-0.330	0.292	0.138	276	<<0.001	0.964	276	<<0.001
		Redeemable Exam	259	15.950	7.214	0.448	-0.351	0.151	-0.406	0.302	0.076	259	0.001	0.960	259	<<0.001
	2015	Lecture Test 1	231	9.100	3.024	0.199	-0.164	0.160	-0.660	0.319	0.103	231	<<0.001	0.974	231	<<0.001
		Lecture Test 2	198	9.370	2.926	0.208	-0.121	0.173	-0.722	0.344	0.095	198	<<0.001	0.976	198	0.002
		Exam Test	300	5.240	1.986	0.115	0.126	0.141	-0.588	0.281	0.118	300	<<0.001	0.969	300	<<0.001
		Redeemable Exam	277	16.970	7.027	0.422	-0.533	0.146	-0.357	0.292	0.107	277	<<0.001	0.956	277	<<0.001

7.2 MCQ Assessment Histograms and Q-Q Plots

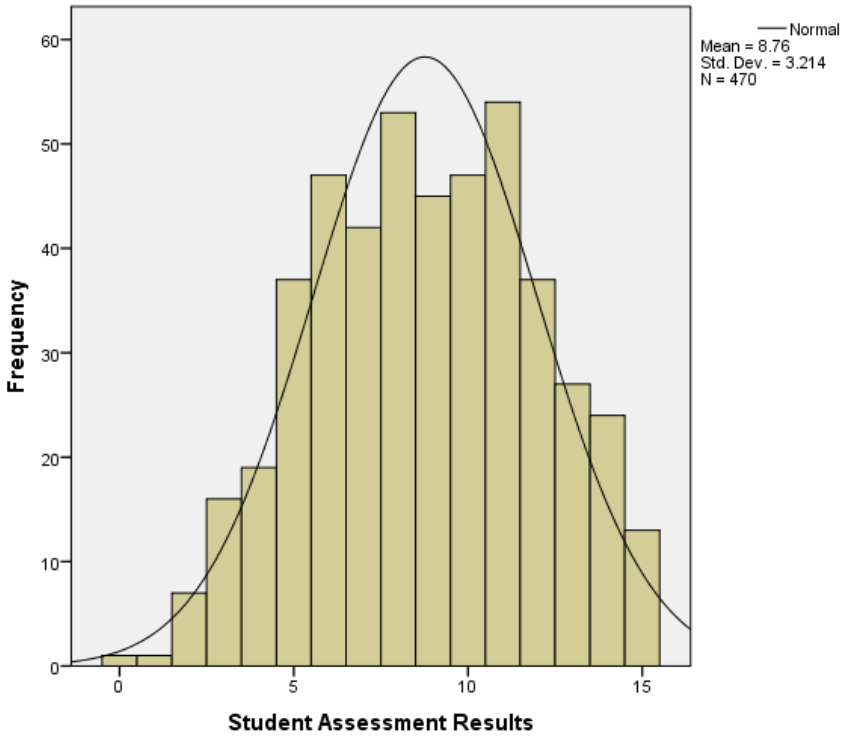

Figure 45: Student Scores Obtained in Chemistry IA Lecture Test 1 2012

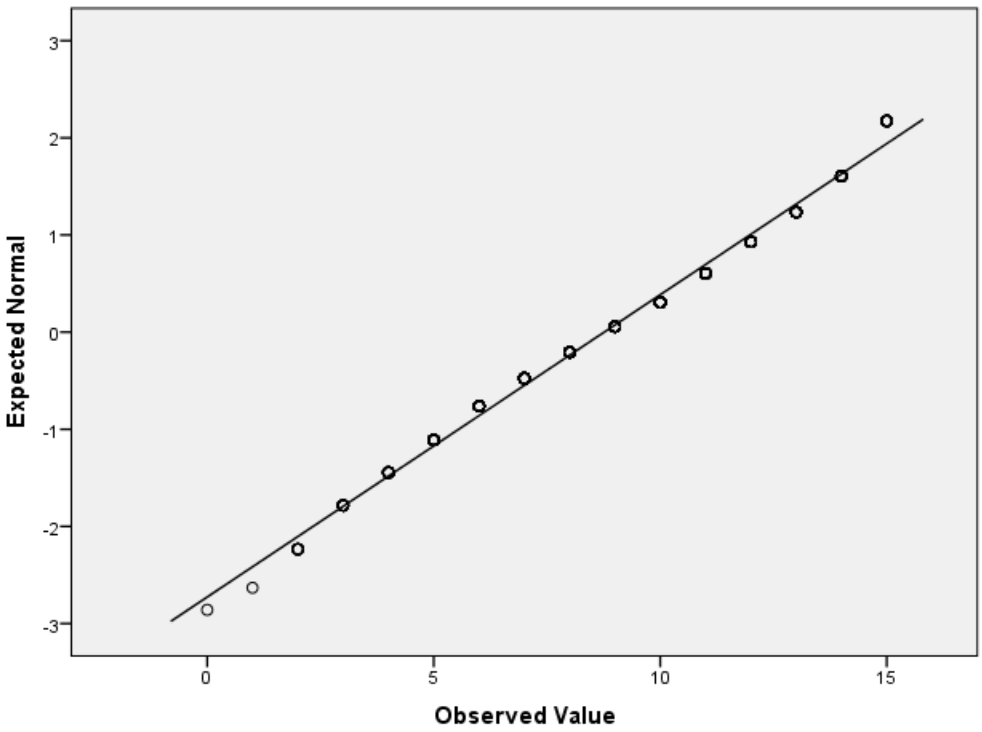

Figure 46: Q-Q Plot of Student Results in Chemistry IA Lecture Test 1 2012

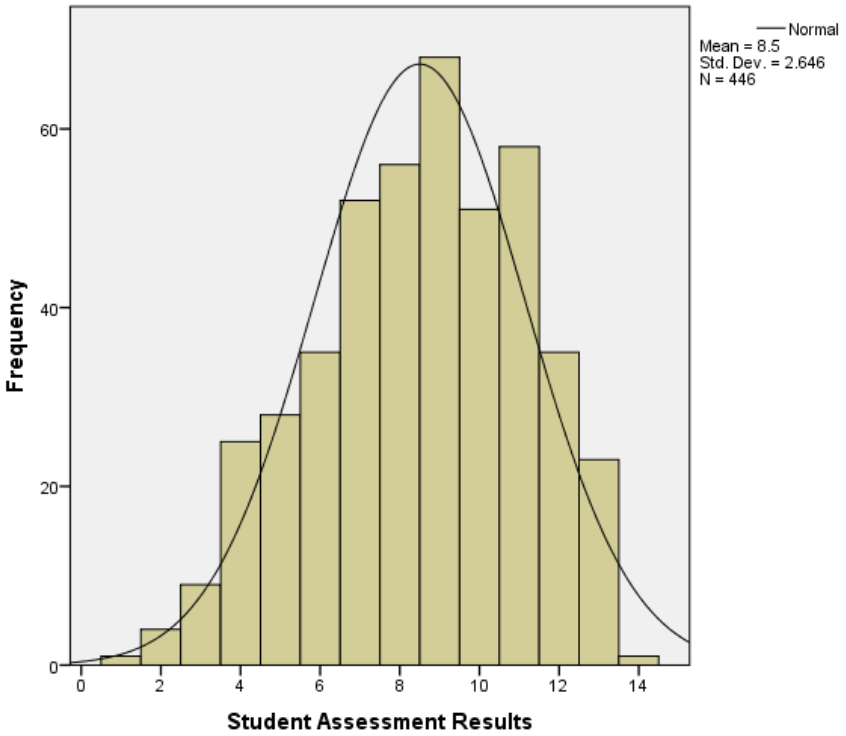

Figure 47: Student Scores Obtained in Chemistry IA Lecture Test 2 2012

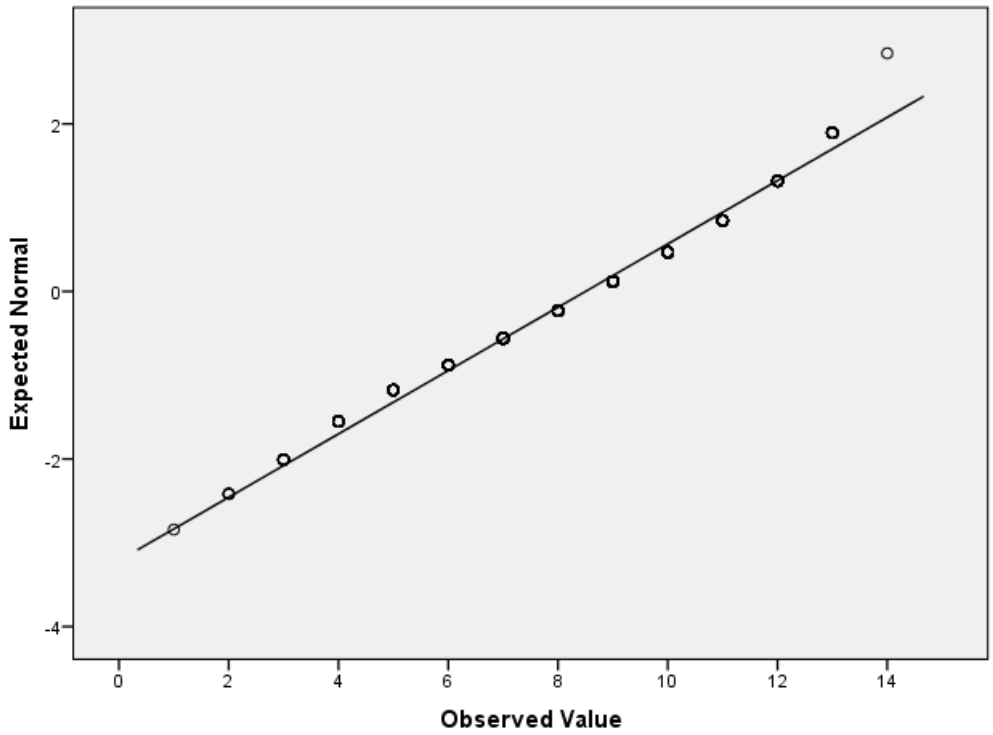

Figure 48: Q-Q Plot of Student Results in Chemistry IA Lecture Test 2 2012

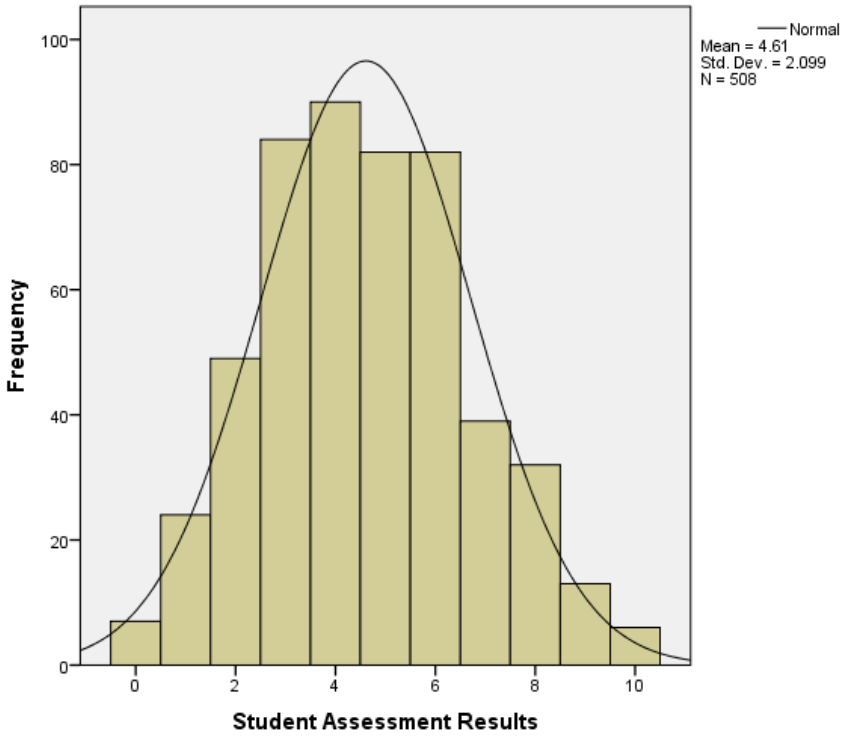

Figure 49: Student Scores Obtained in Chemistry IA Exam 2012

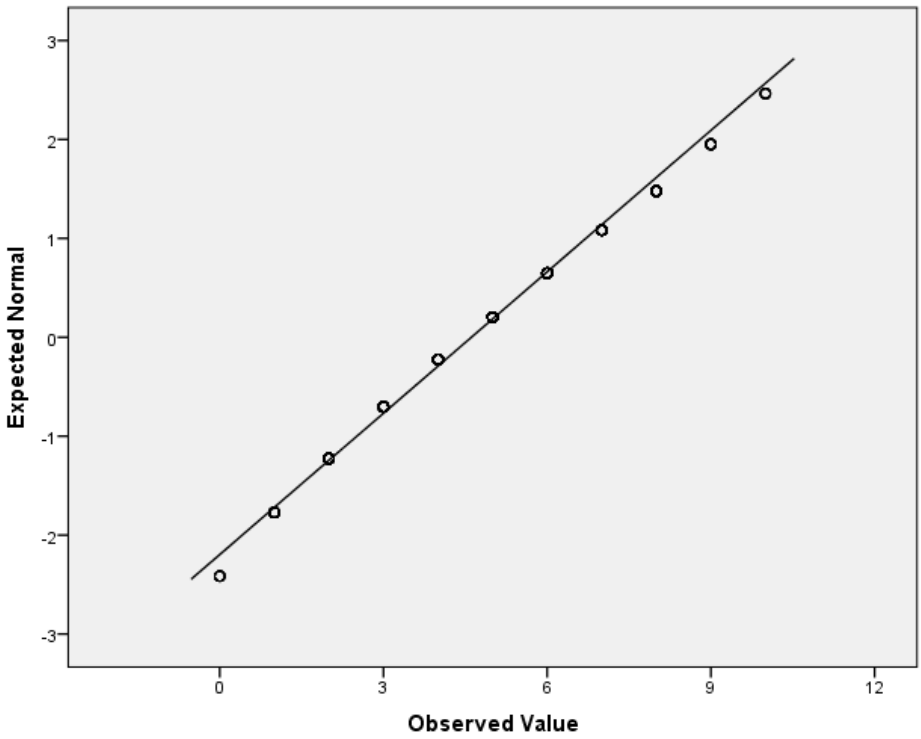

Figure 50: Q-Q Plot of Student Results in Chemistry IA Exam 2012

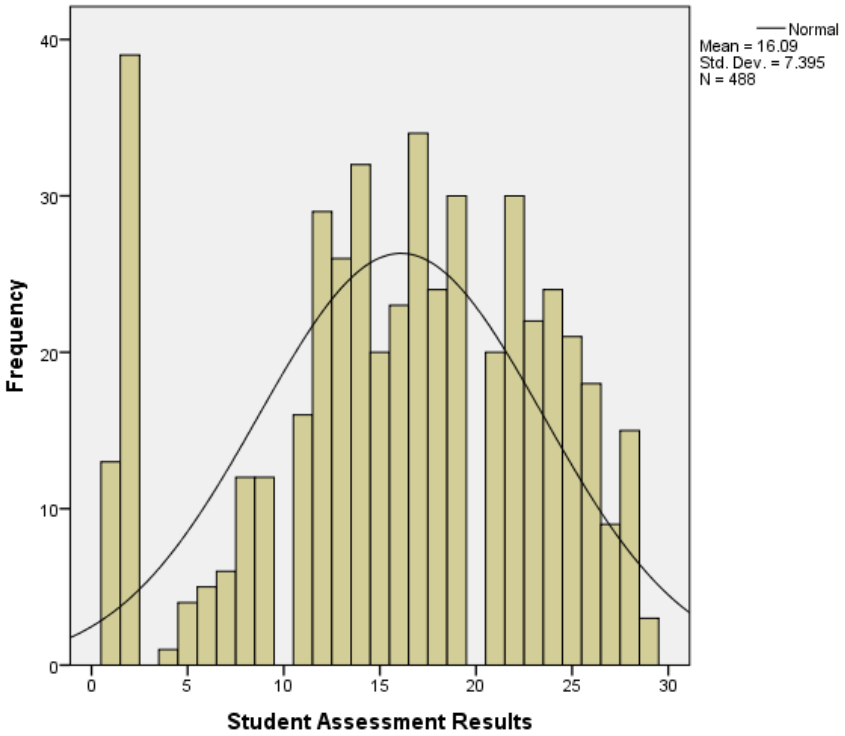

Figure 51: Student Scores Obtained in Chemistry IA Redeemable Exam 2012

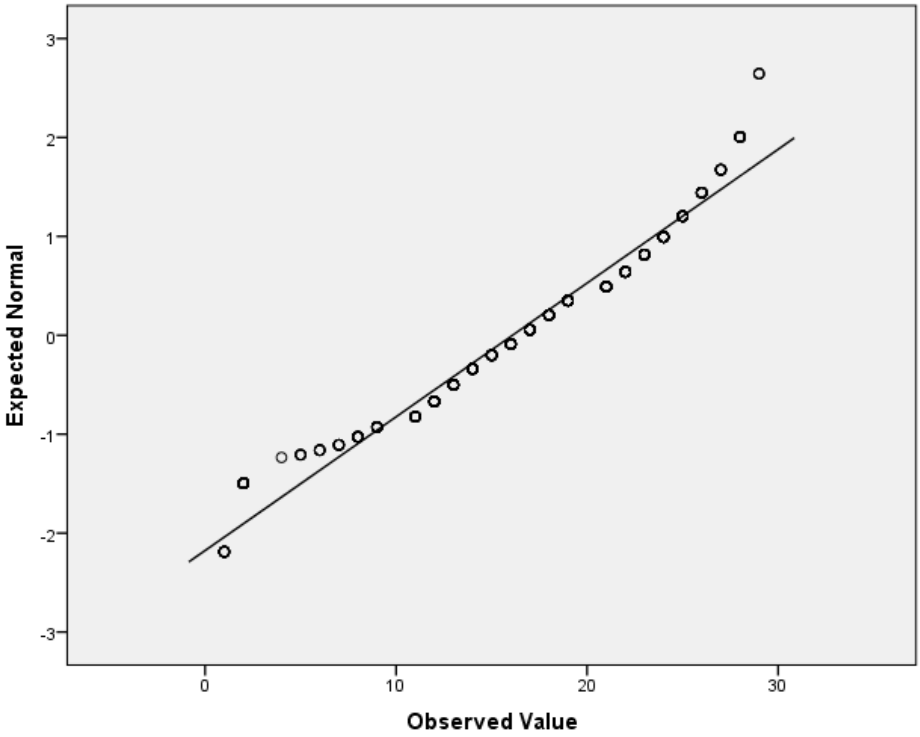

Figure 52: Q-Q Plot of Student Results in Chemistry IA Redeemable Exam 2012

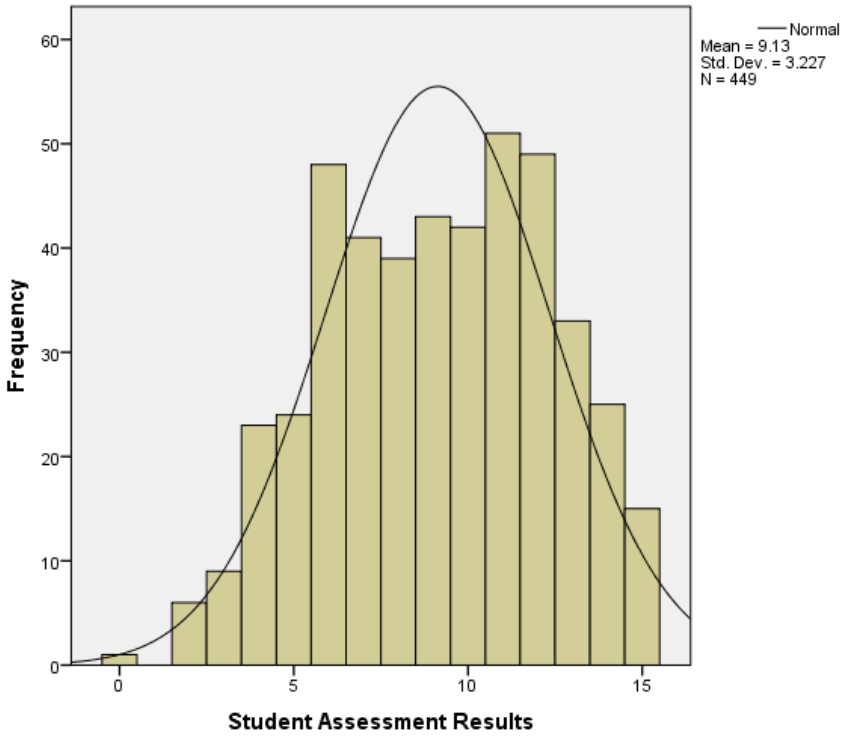

Figure 53: Student Scores Obtained in Chemistry IA Lecture Test 1 2013

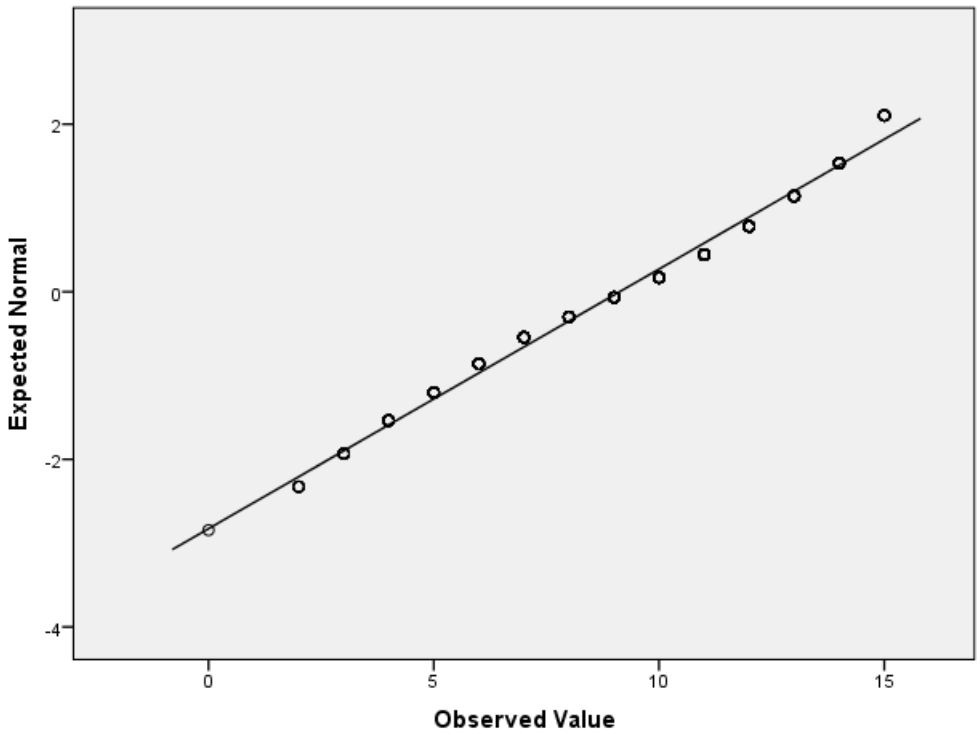

Figure 54: Q-Q Plot of Student Results in Chemistry IA Lecture Test 1 2013

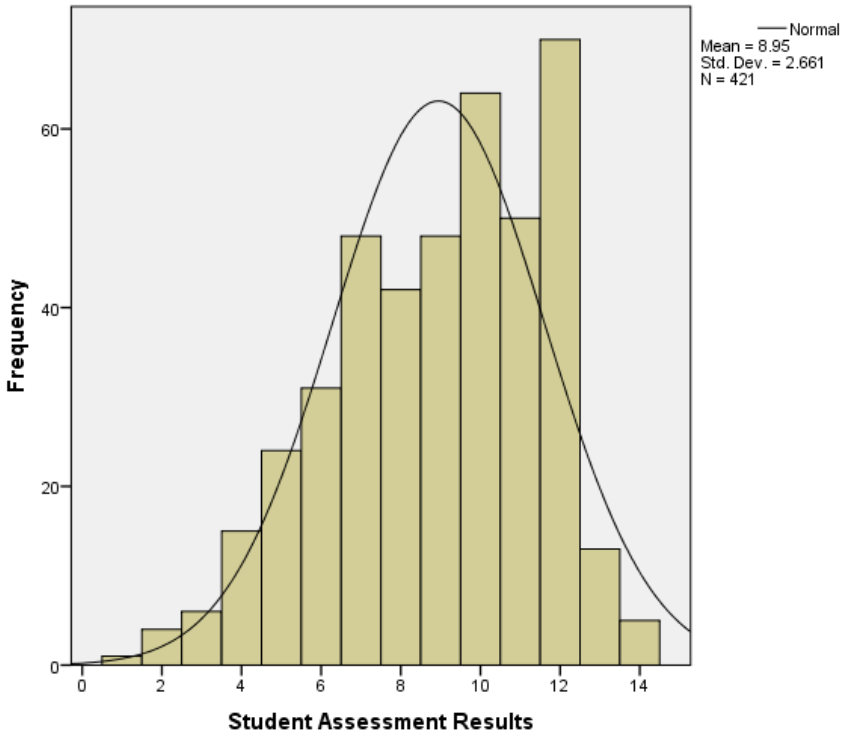

Figure 55: Student Scores Obtained in Chemistry IA Lecture Test 2 2013

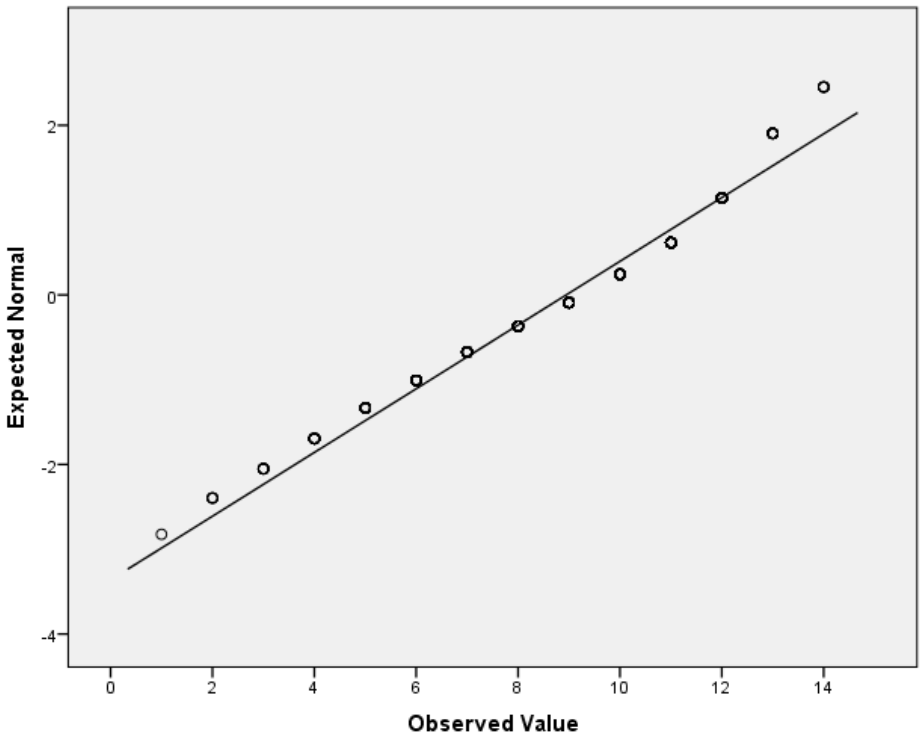

Figure 56: Q-Q Plot of Student Results in Chemistry IA Lecture Test 2 2013

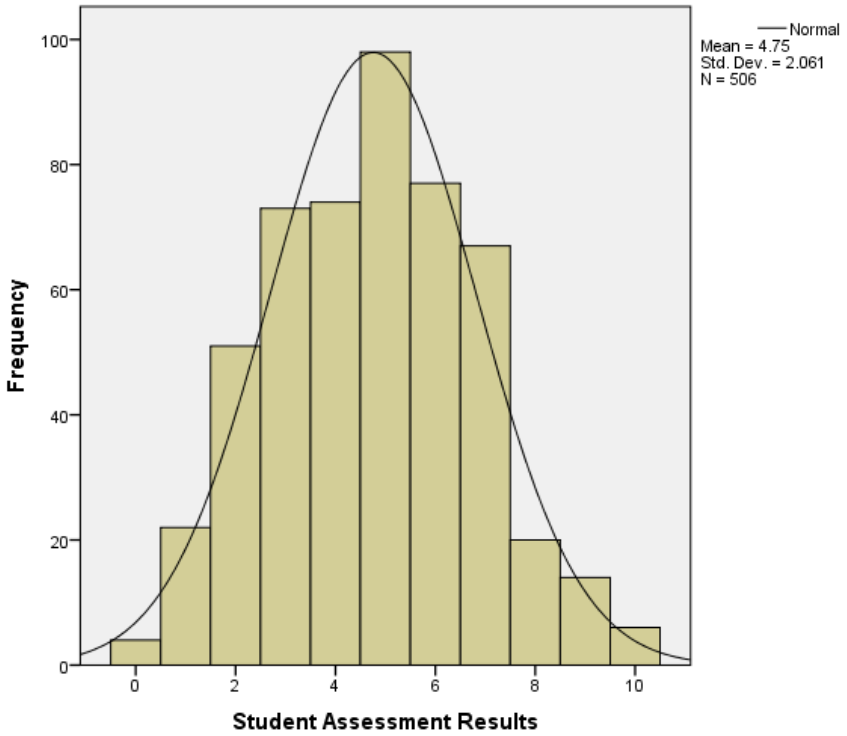

Figure 57: Student Scores Obtained in Chemistry IA Exam 2013

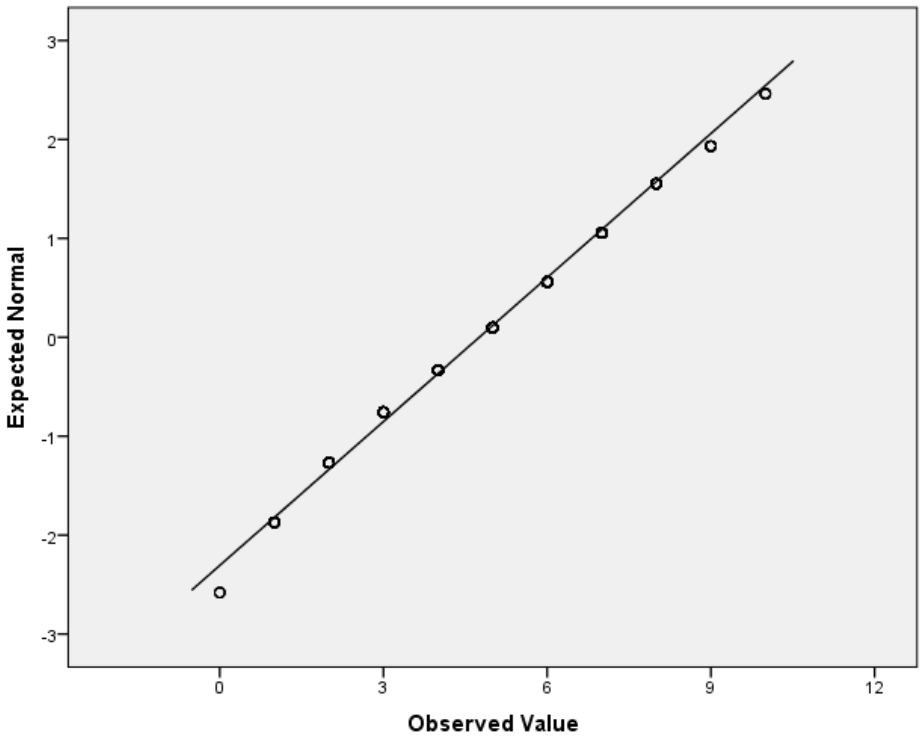

Figure 58: Q-Q Plot of Student Results in Chemistry IA Exam 2013

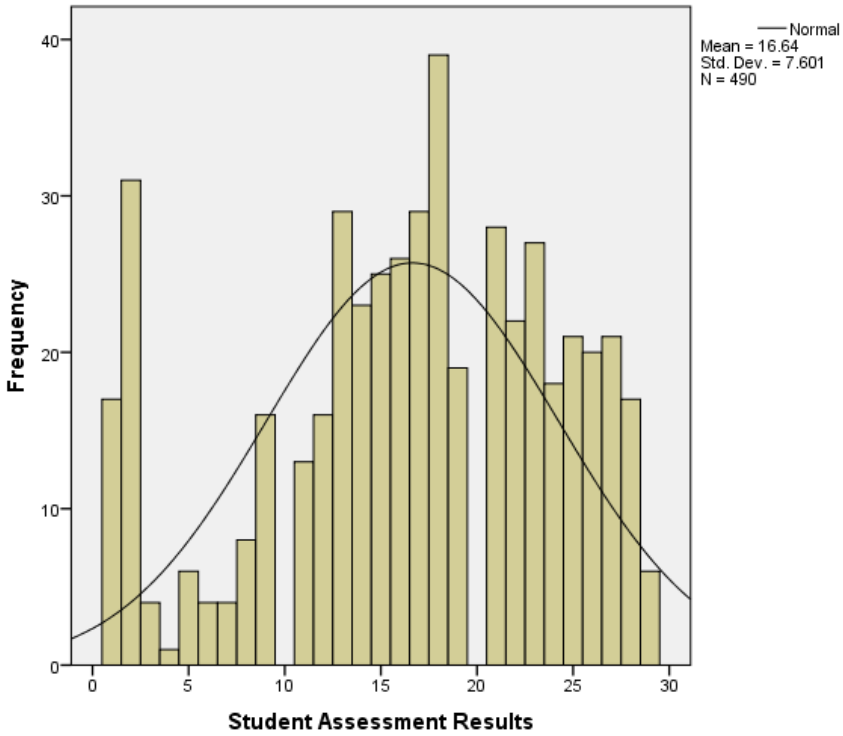

Figure 59: Student Scores Obtained in Chemistry IA Redeemable Exam 2013

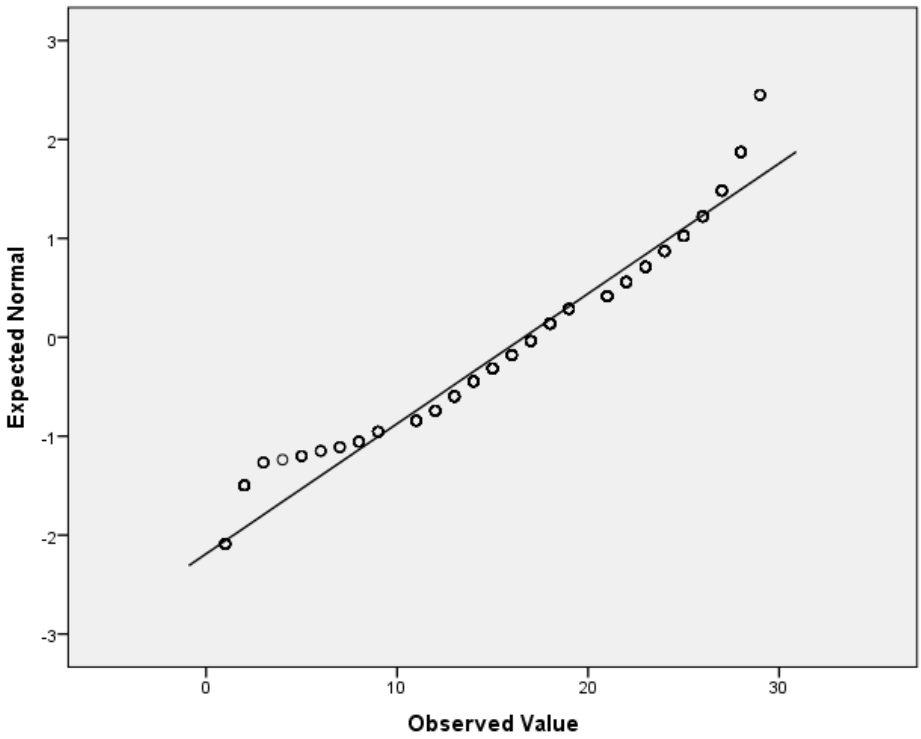

Figure 60: Q-Q Plot of Student Results in Chemistry IA Redeemable Exam 2013

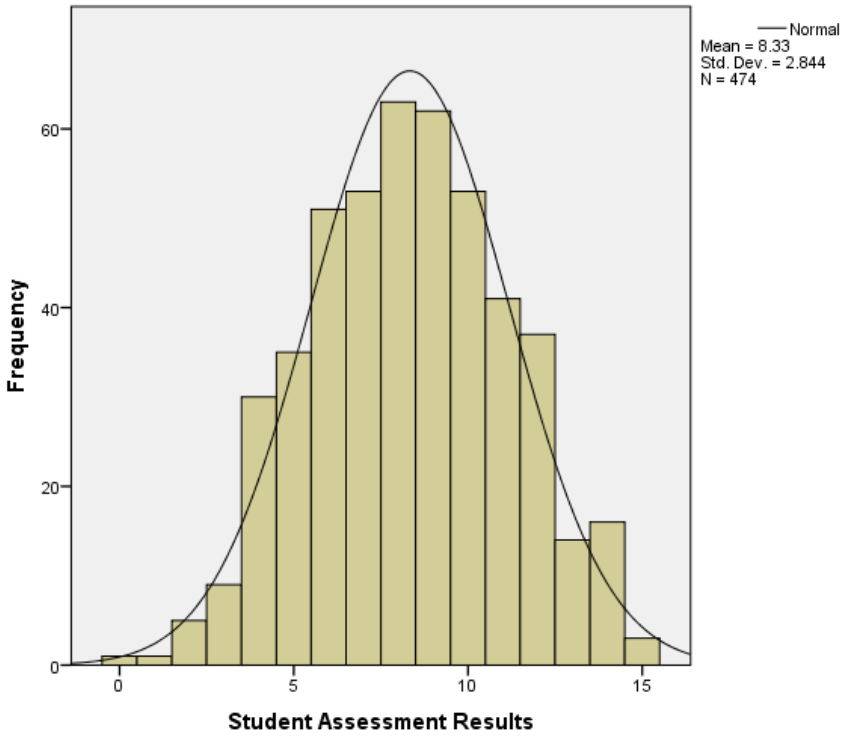

Figure 61: Student Scores Obtained in Chemistry IA Lecture Test 1 2014

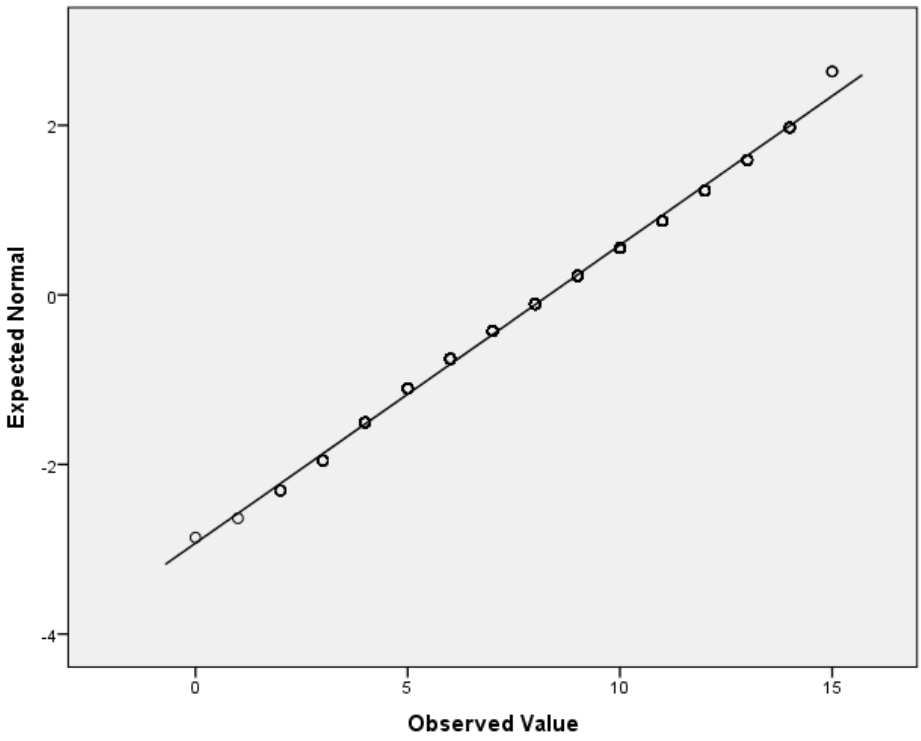

Figure 62: Q-Q Plot of Student Results in Chemistry IA Lecture Test 1 2014

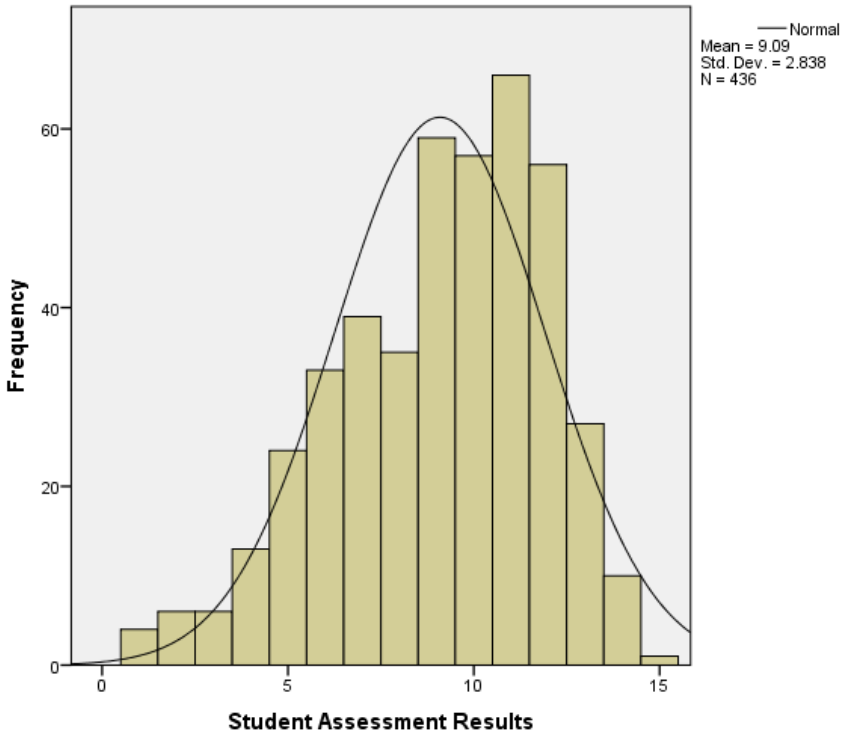

Figure 63: Student Scores Obtained in Chemistry IA Lecture Test 2 2014

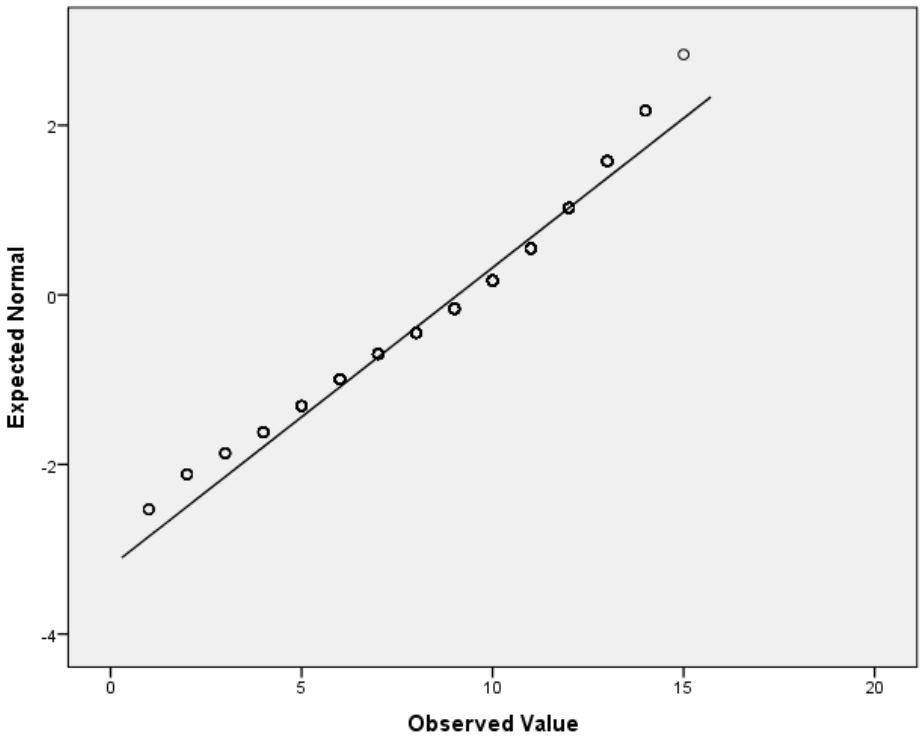

Figure 64: Q-Q Plot of Student Results in Chemistry IA Lecture Test 2 2014

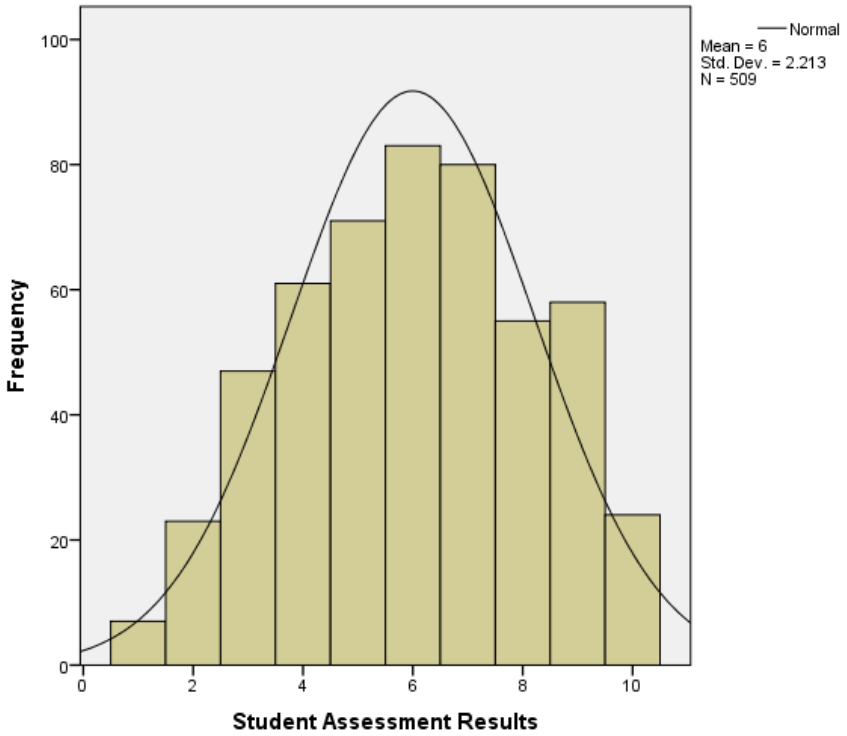

Figure 65: Student Scores Obtained in Chemistry IA Exam 2014

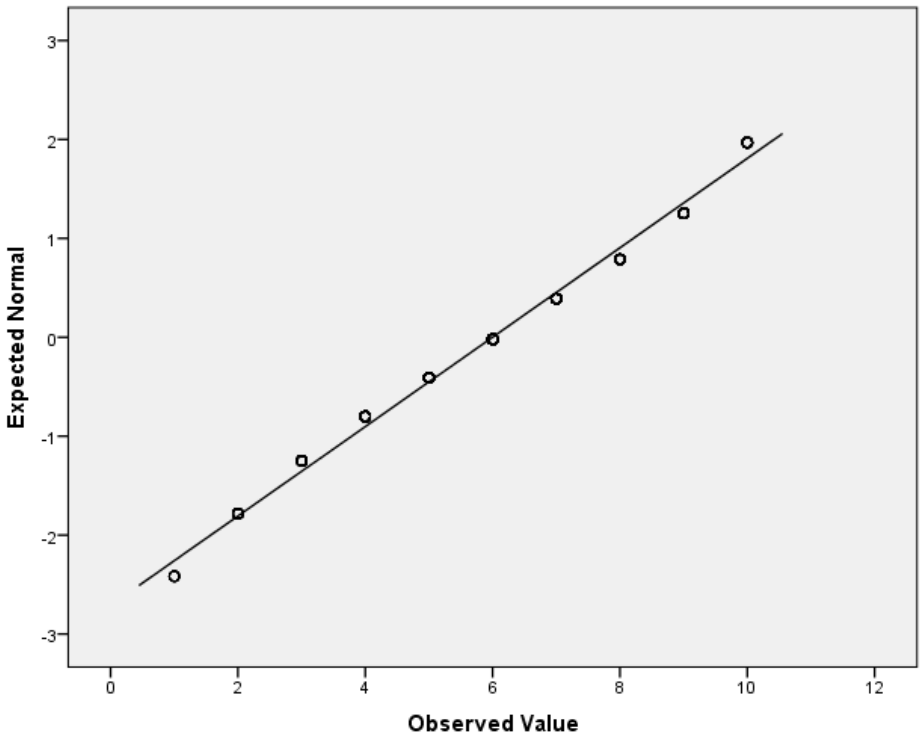

Figure 66: Q-Q Plot of Student Results in Chemistry IA Exam 2014

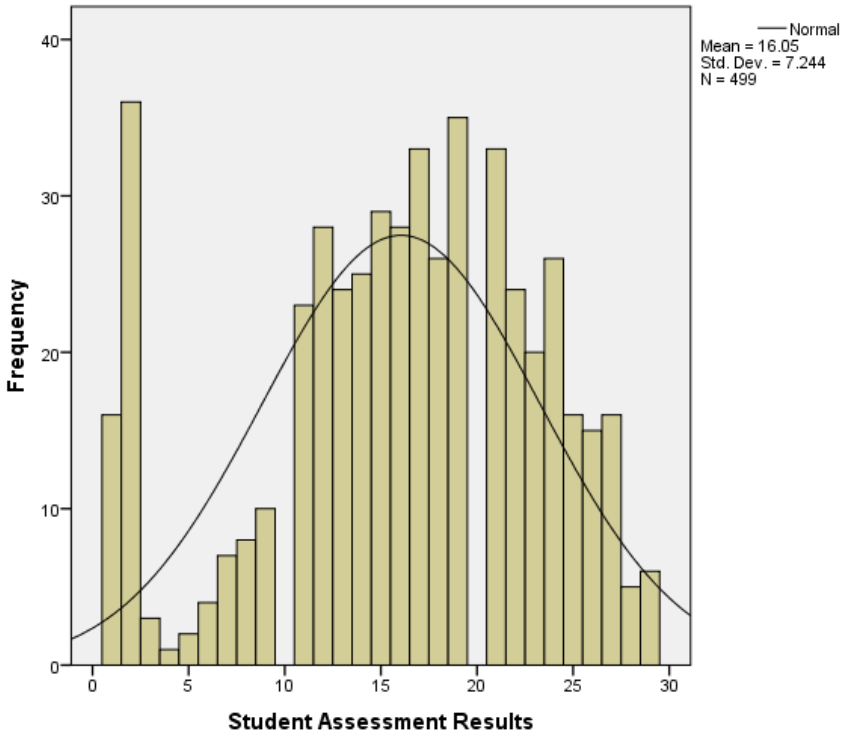

Figure 67: Student Scores Obtained in Chemistry IA Redeemable Exam 2014

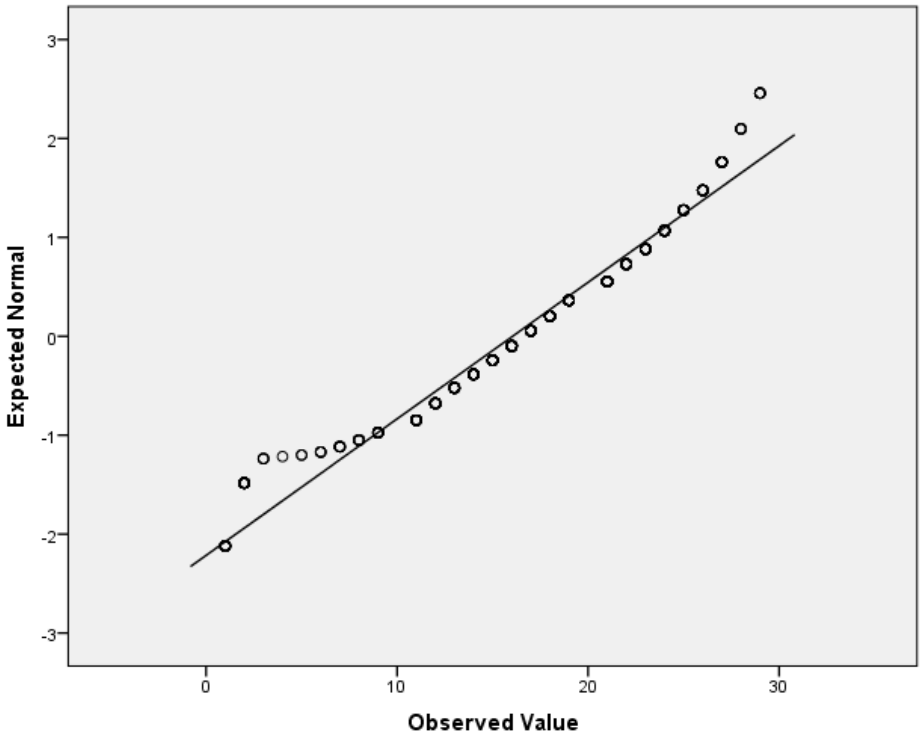

Figure 68: Q-Q Plot of Student Results in Chemistry IA Redeemable Exam

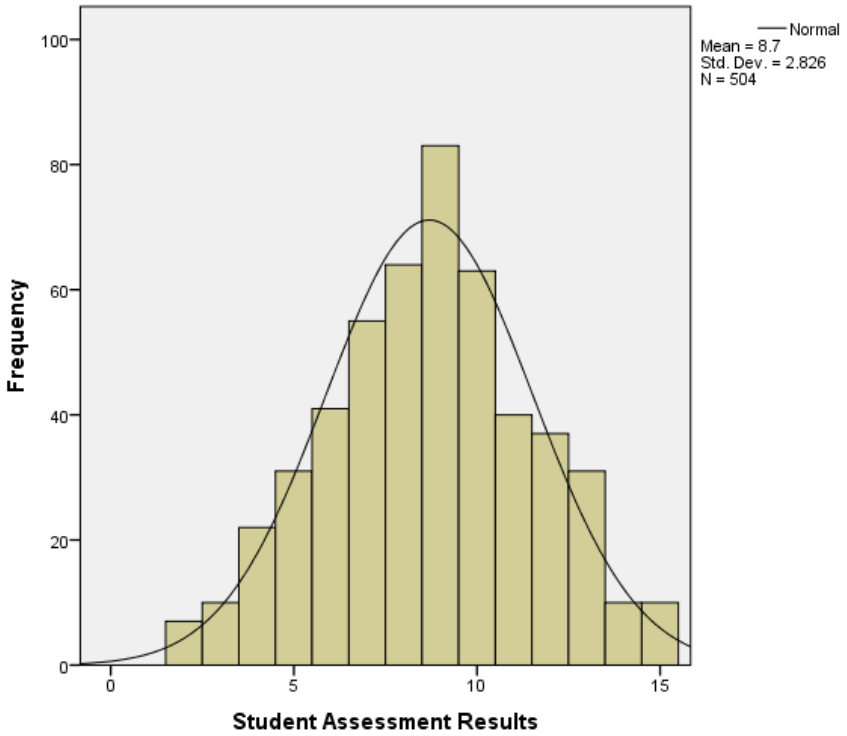

Figure 69: Student Scores Obtained in Chemistry IA Lecture Test 1 2015

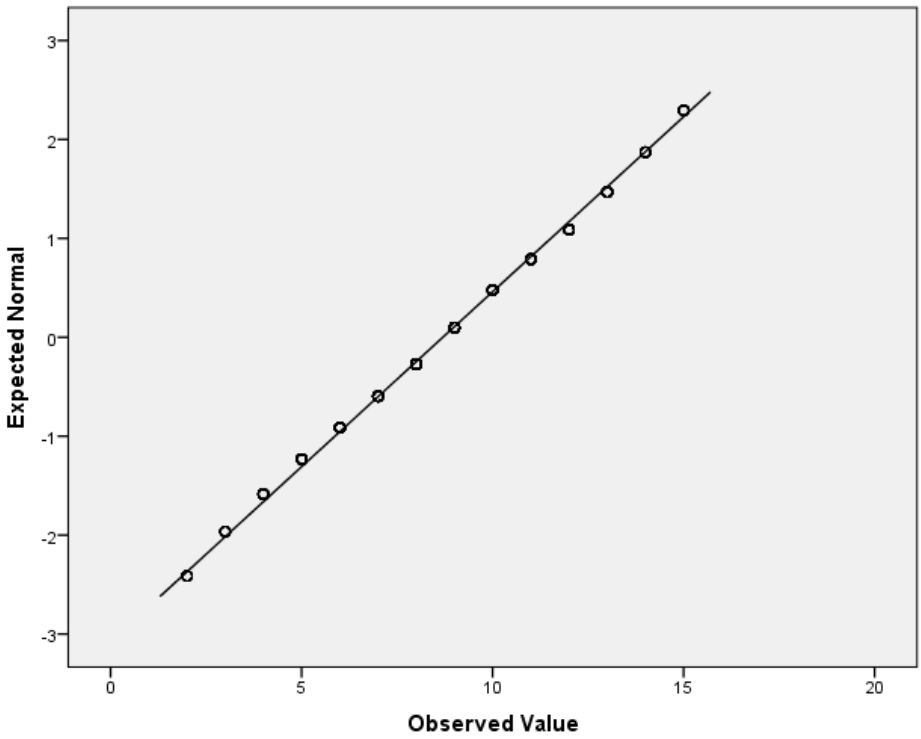

Figure 70: Q-Q Plot of Student Results in Chemistry IA Lecture Test 1 2015

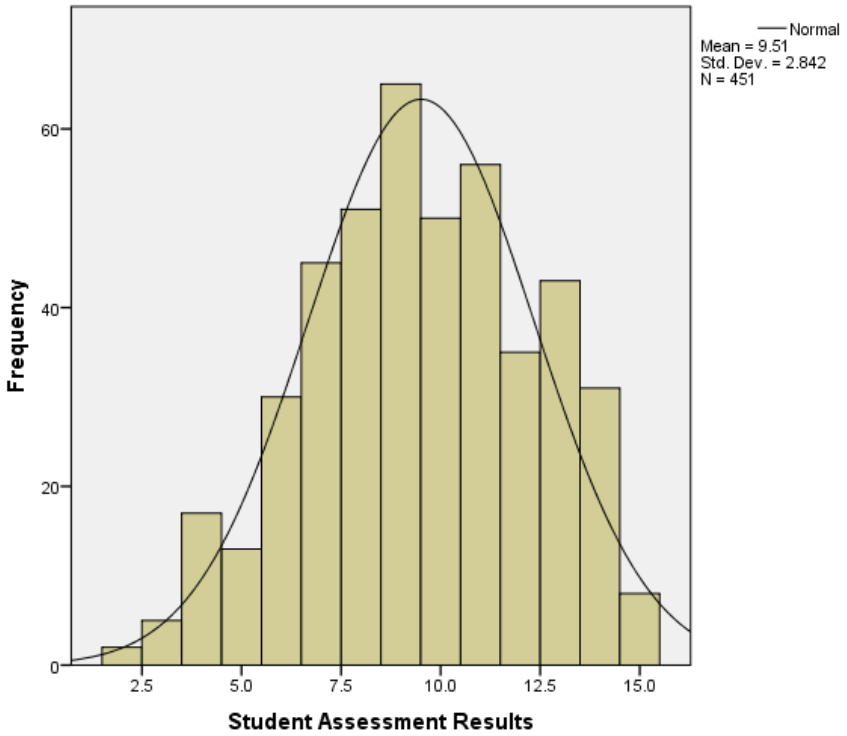

Figure 71: Student Scores Obtained in Chemistry IA Lecture Test 2 2015

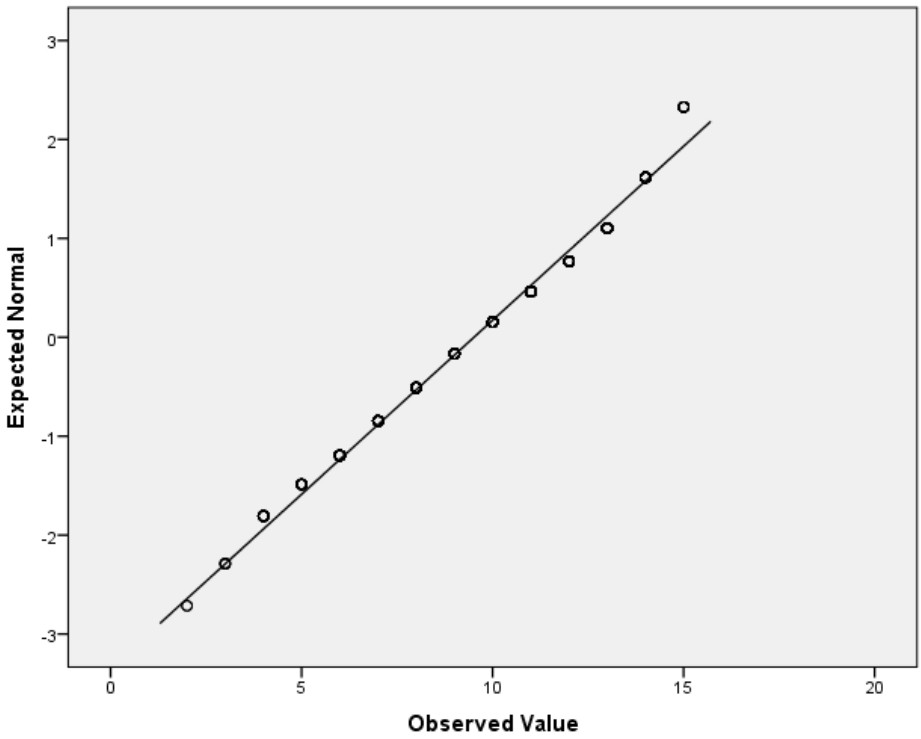

Figure 72: Q-Q Plot of Student Results in Chemistry IA Lecture Test 2 2015

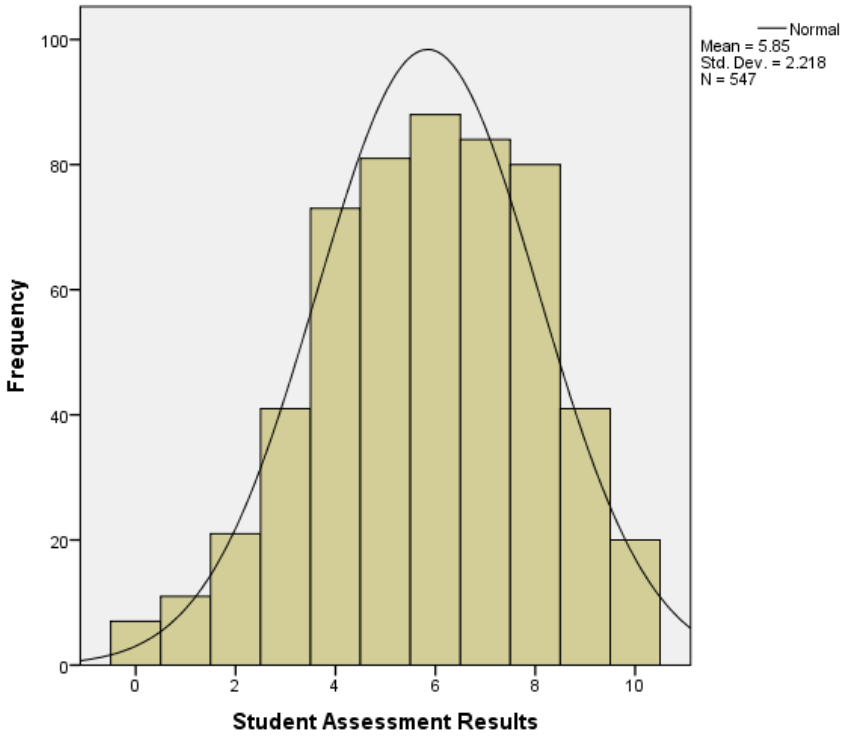

Figure 73: Student Scores Obtained in Chemistry IA Exam 2015

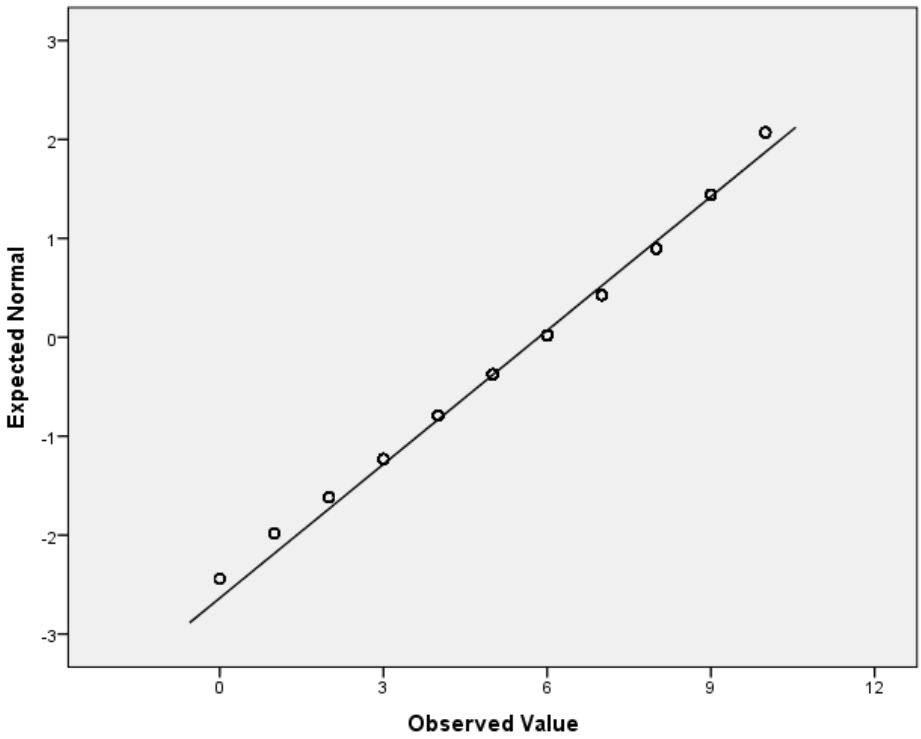

Figure 74: Q-Q Plot of Student Results in Chemistry IA Exam 2015

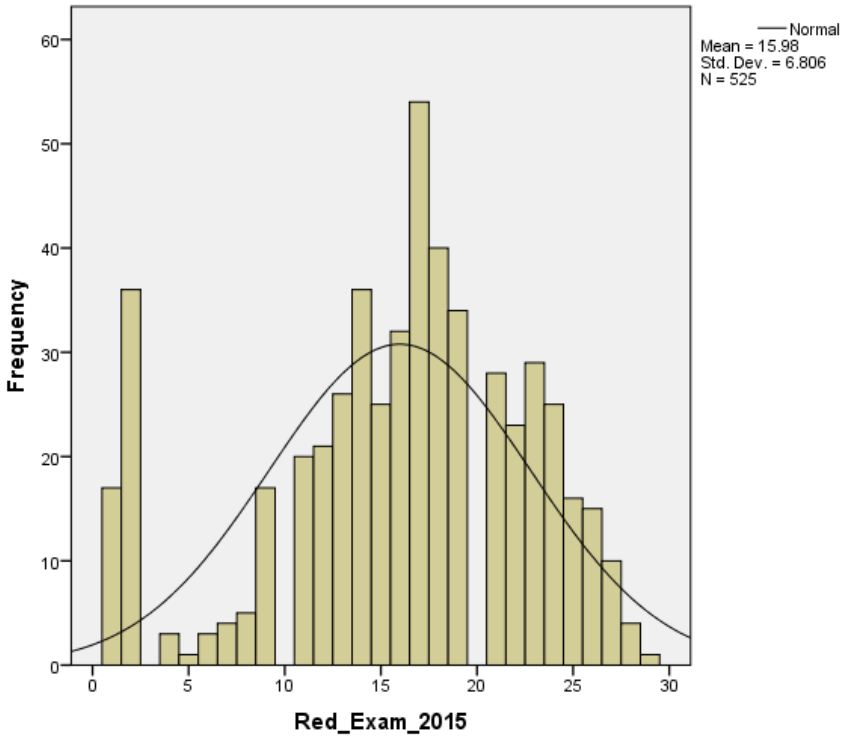

Figure 75: Student Scores Obtained in Chemistry IA Redeemable Exam 2015

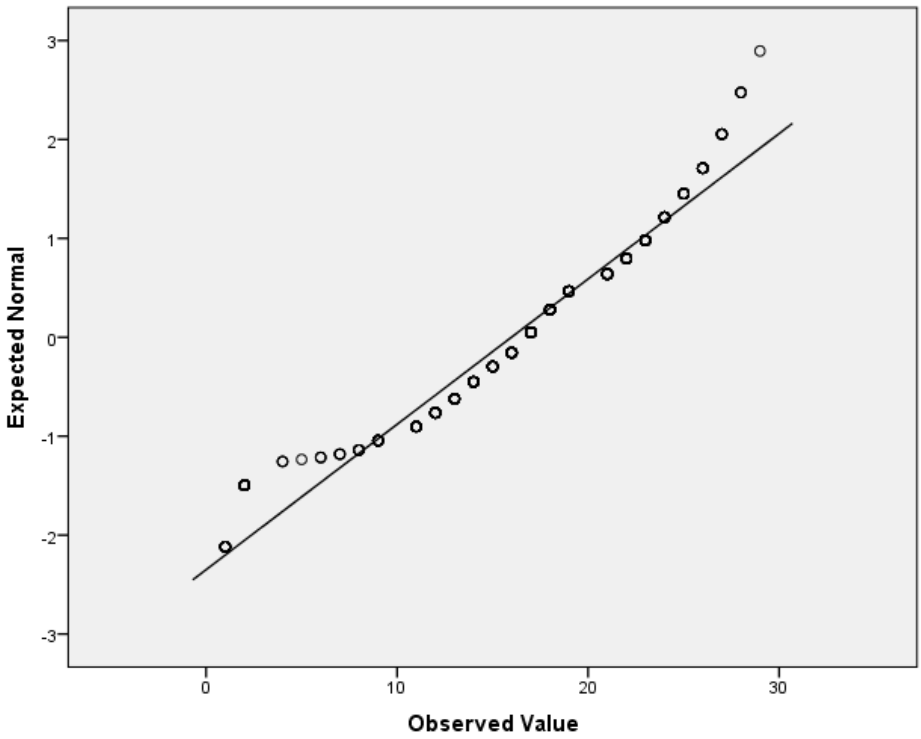

Figure 76: Q-Q Plot of Student Results in Chemistry IA Redeemable Exam 2015

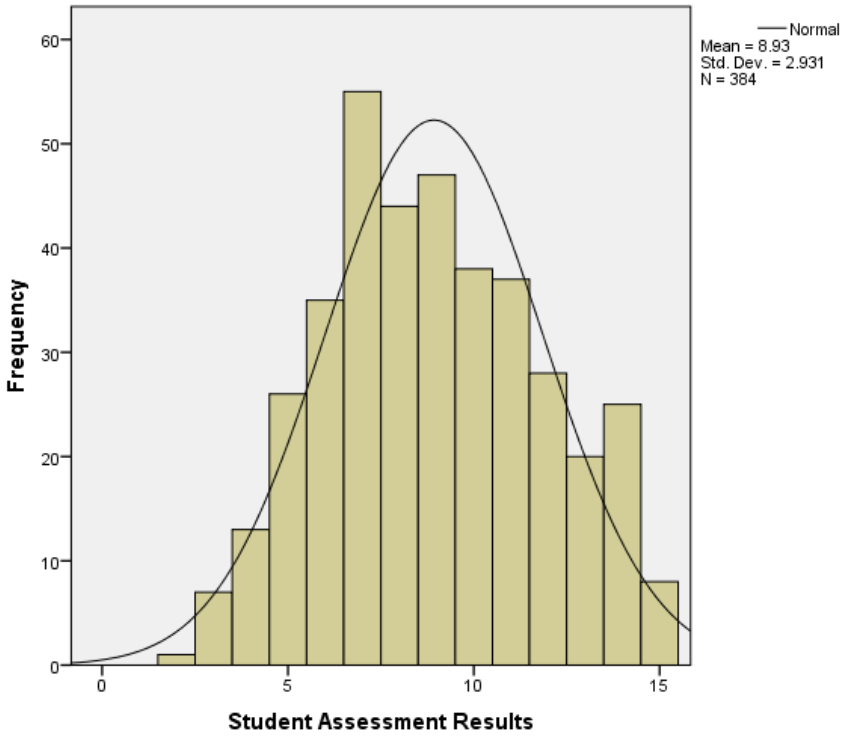

Figure 77: Student Scores Obtained in Chemistry IB Lecture Test 1 2012

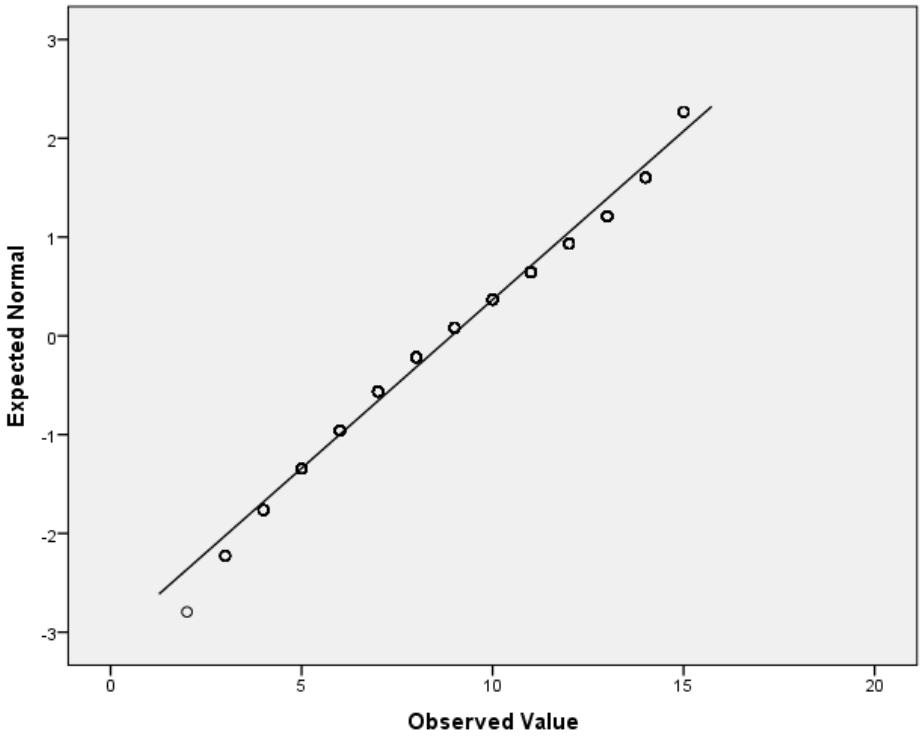

Figure 78: Q-Q Plot of Student Results in Chemistry IB Lecture Test 1 2012

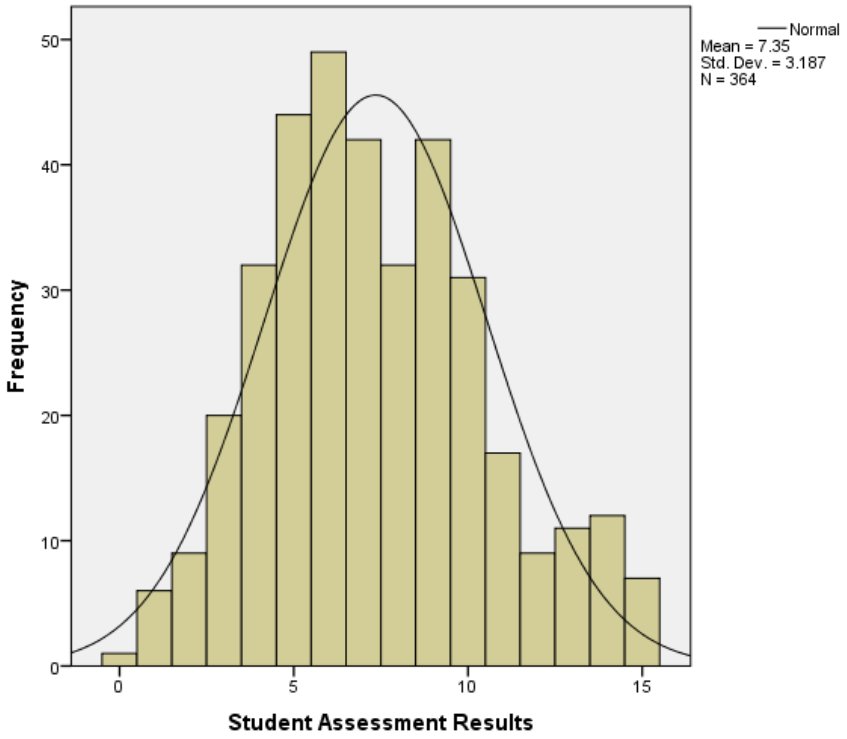

Figure 79: Student Scores Obtained in Chemistry IB Lecture Test 2 2012

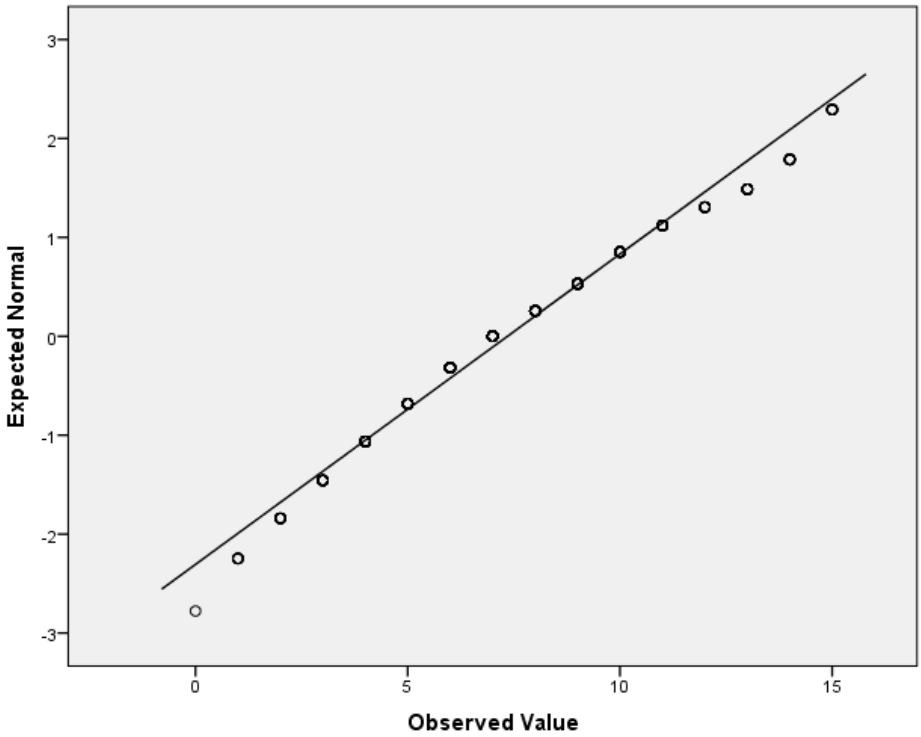

Figure 80: Q-Q Plot of Student Results in Chemistry IB Lecture Test 2 2012

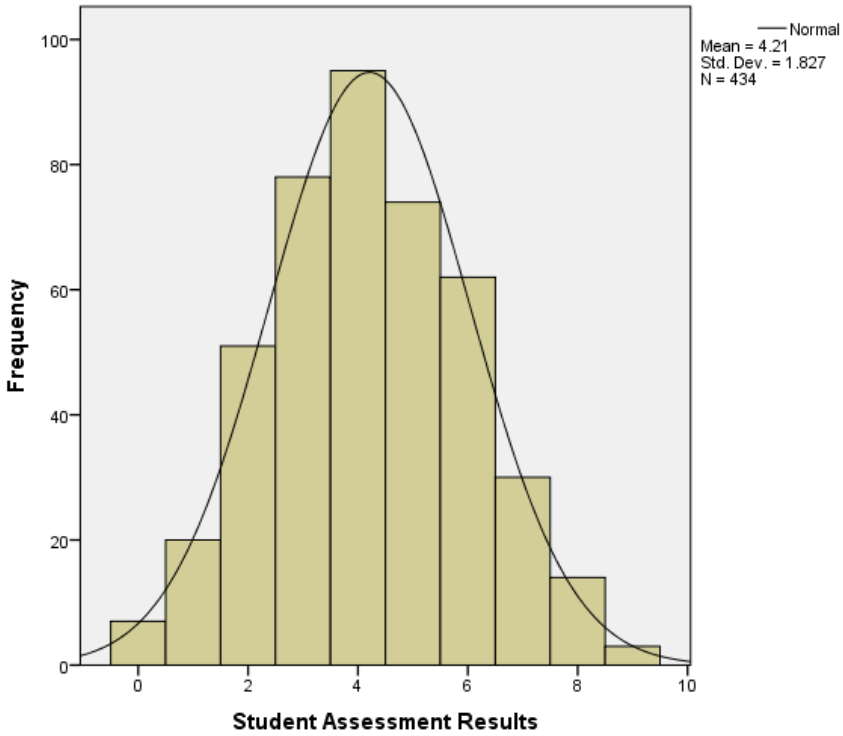

Figure 81: Student Scores Obtained in Chemistry IB Exam 2012

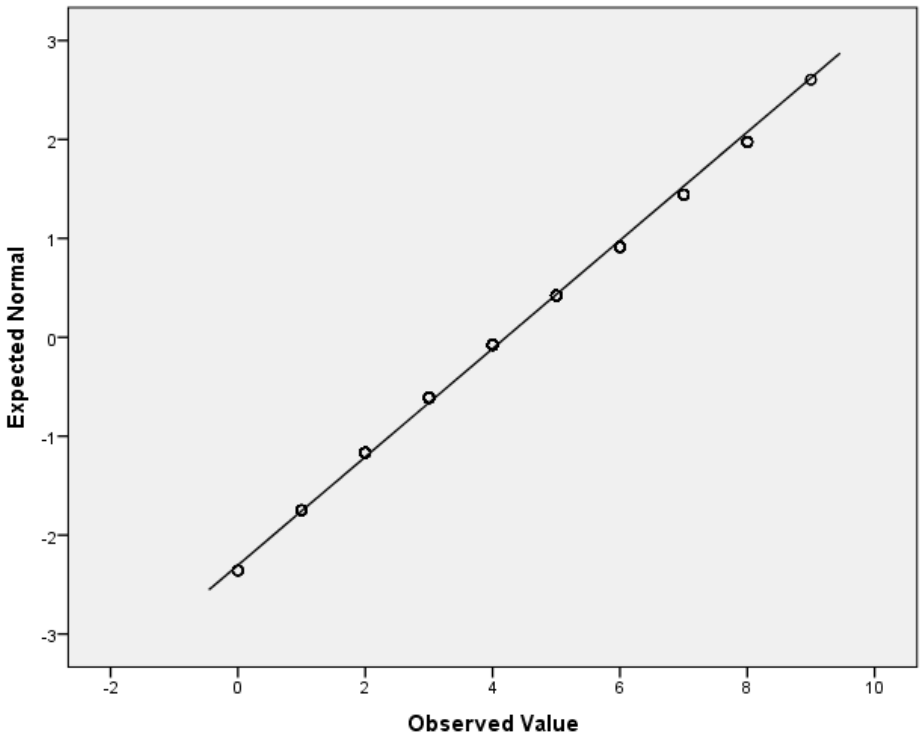

Figure 82: Q-Q Plot of Student Results in Chemistry IB Exam 2012

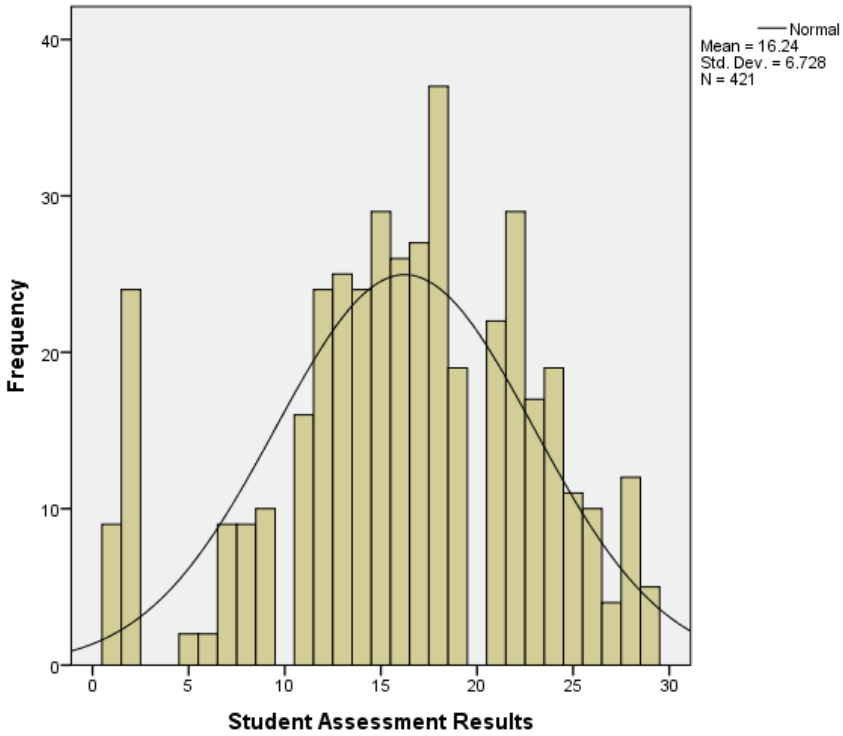

Figure 83: Student Scores Obtained in Chemistry IB Redeemable Exam 2012

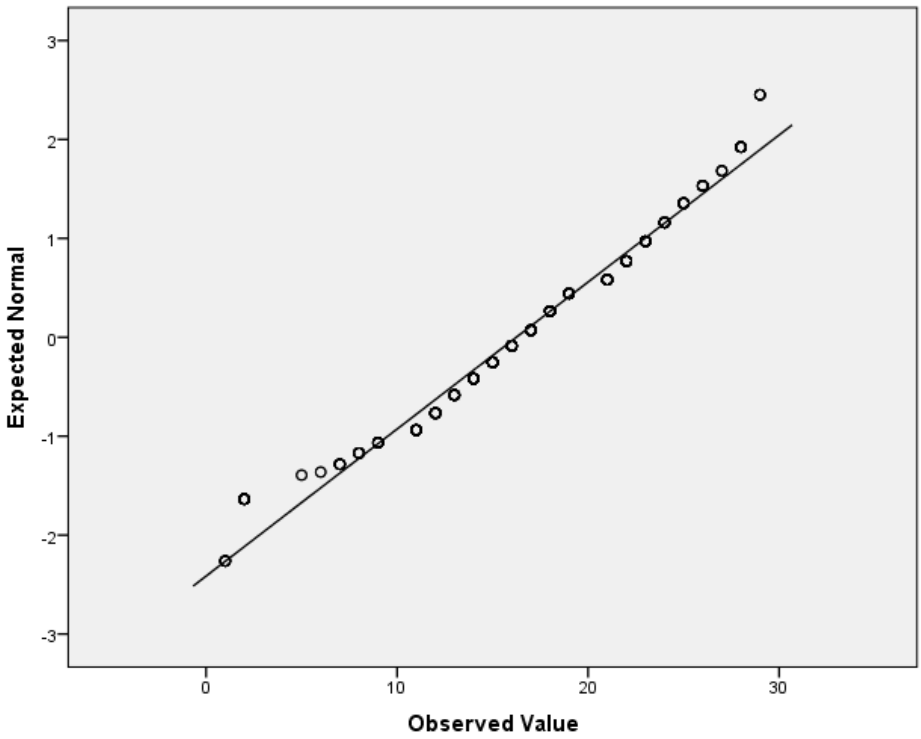

Figure 84: Q-Q Plot of Student Results in Chemistry IB Redeemable Exam 2012

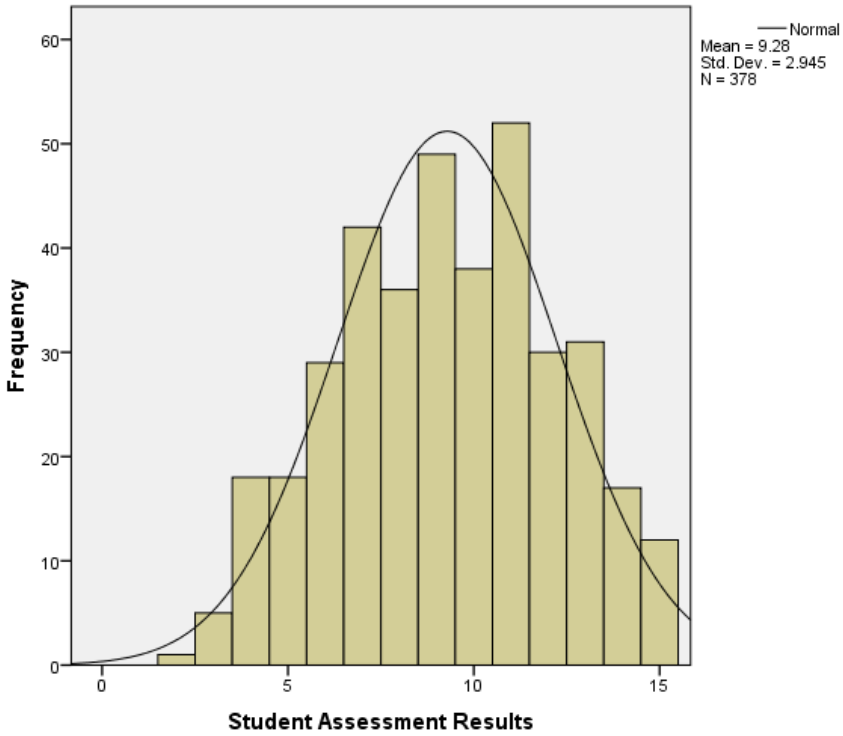

Figure 85: Student Scores Obtained in Chemistry IB Lecture Test 1 2013

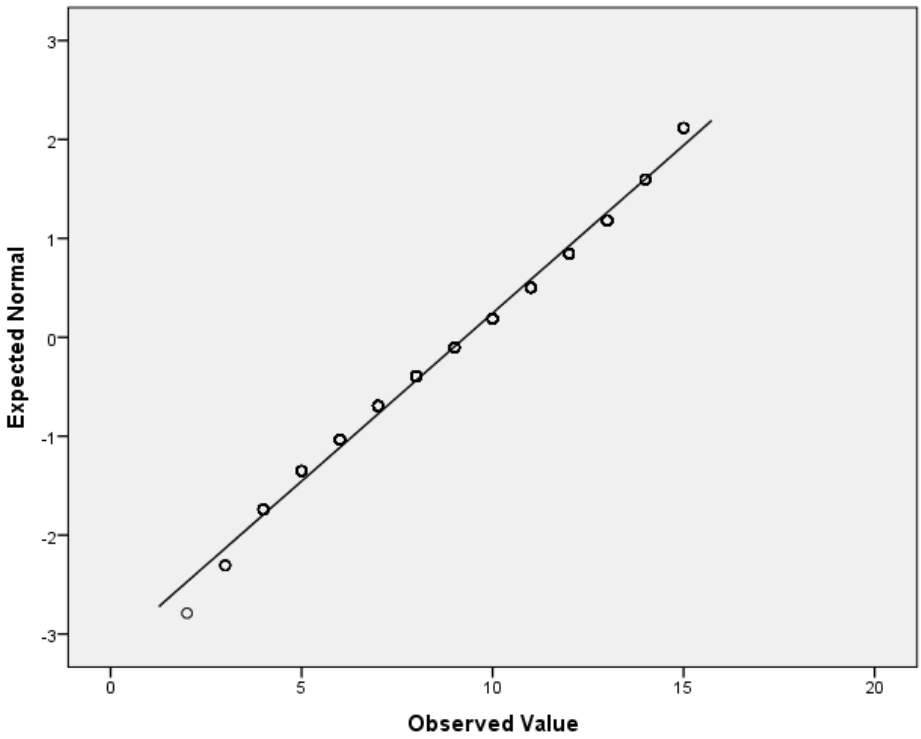

Figure 86: Q-Q Plot of Student Results in Chemistry IB Lecture Test 1 2013

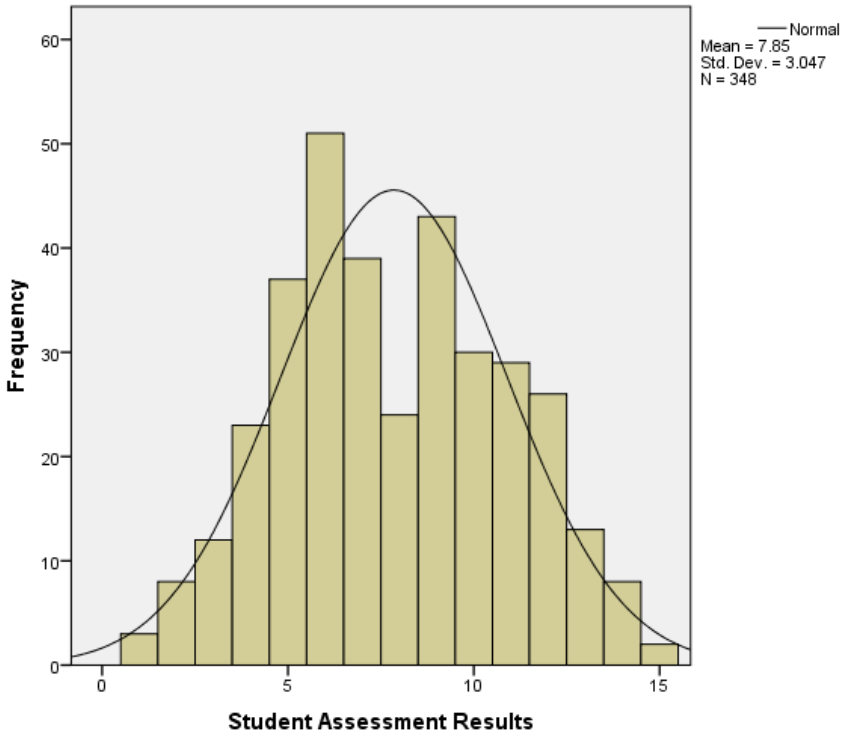

Figure 87: Student Scores Obtained in Chemistry IB Lecture Test 2 2013

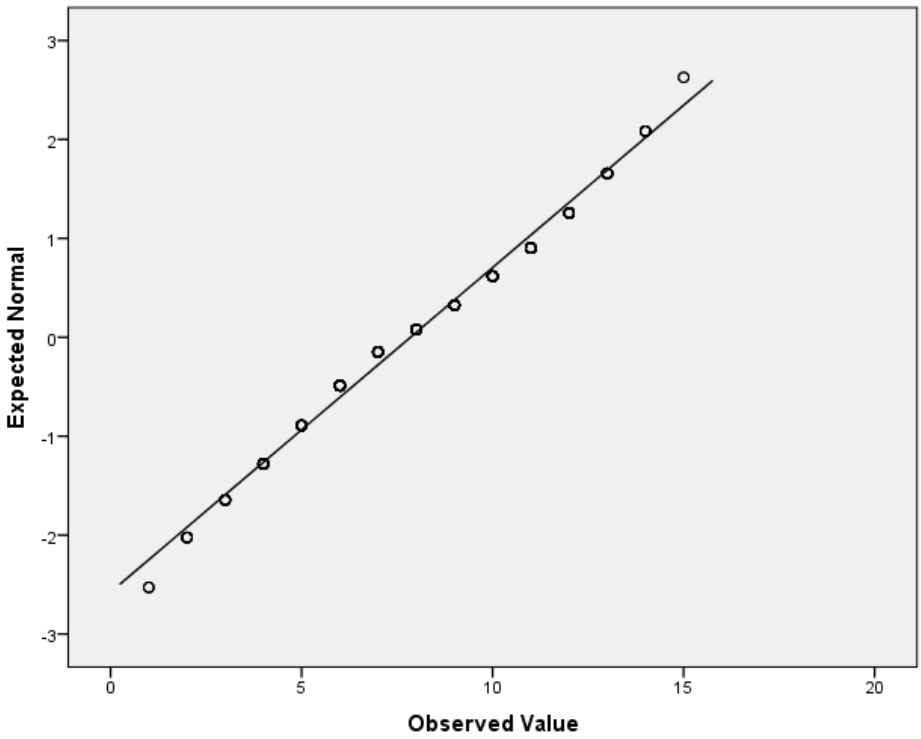

Figure 88: Q-Q Plot of Student Results in Chemistry IB Lecture Test 2 2013

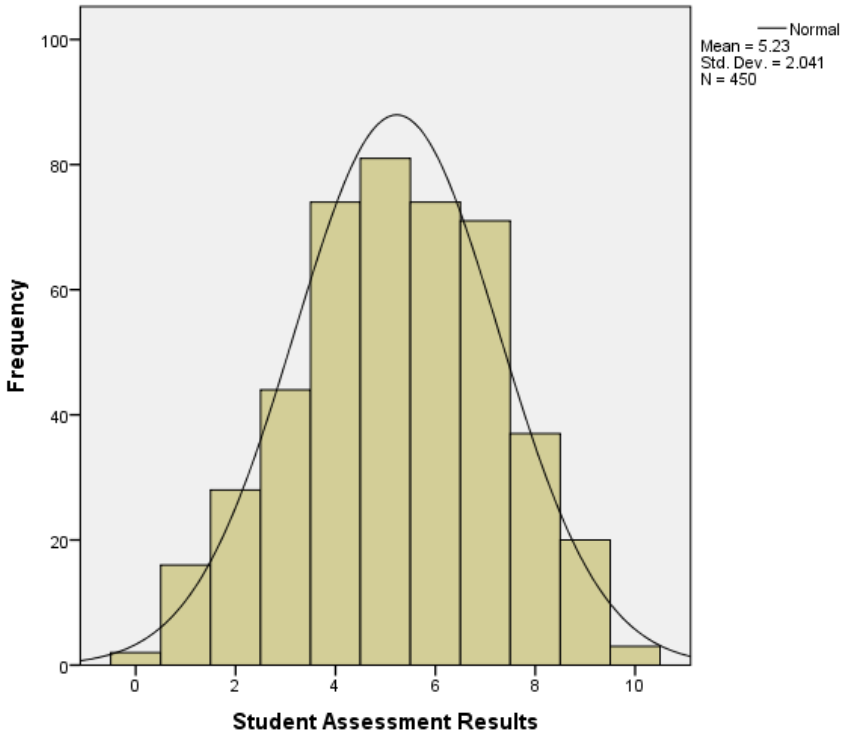

Figure 89: Student Scores Obtained in Chemistry IB Exam 2013

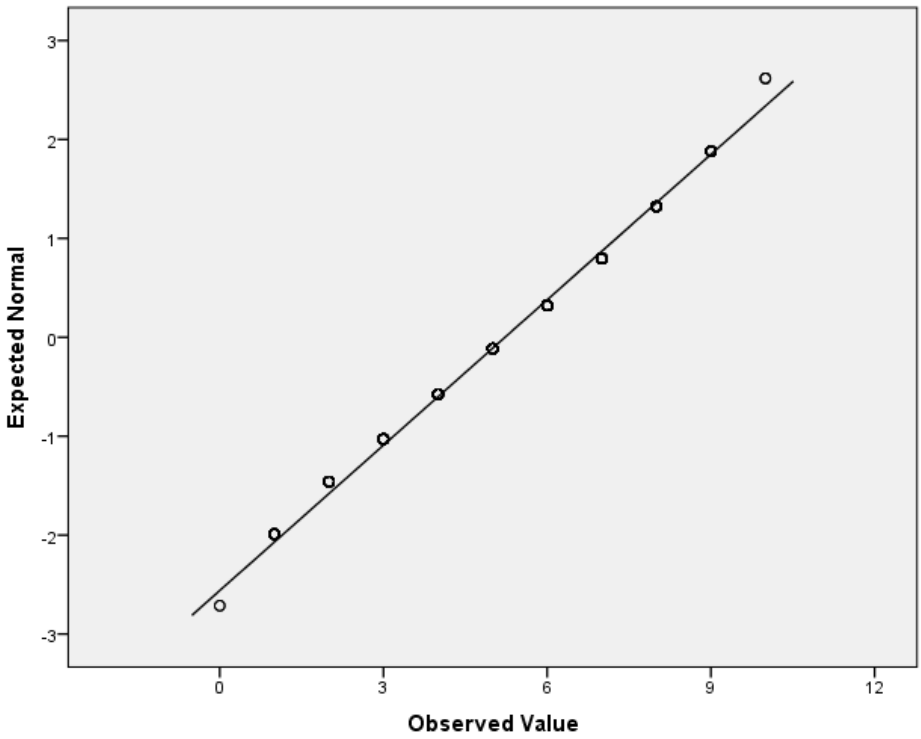

Figure 90: Q-Q Plot of Student Results in Chemistry IB Exam 2013

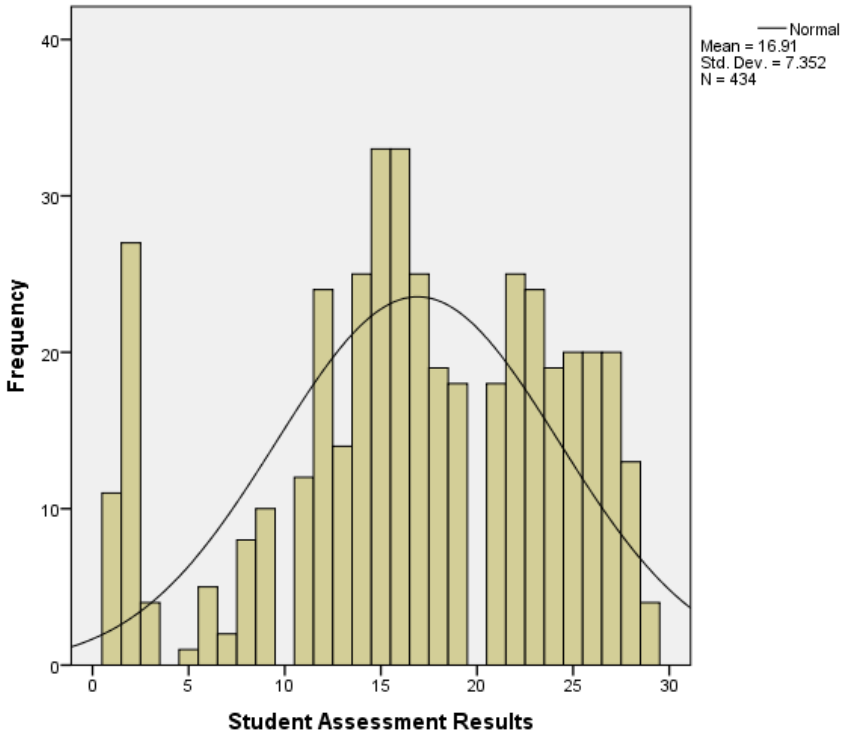

Figure 91: Student Scores Obtained in Chemistry IB Redeemable Exam 2013

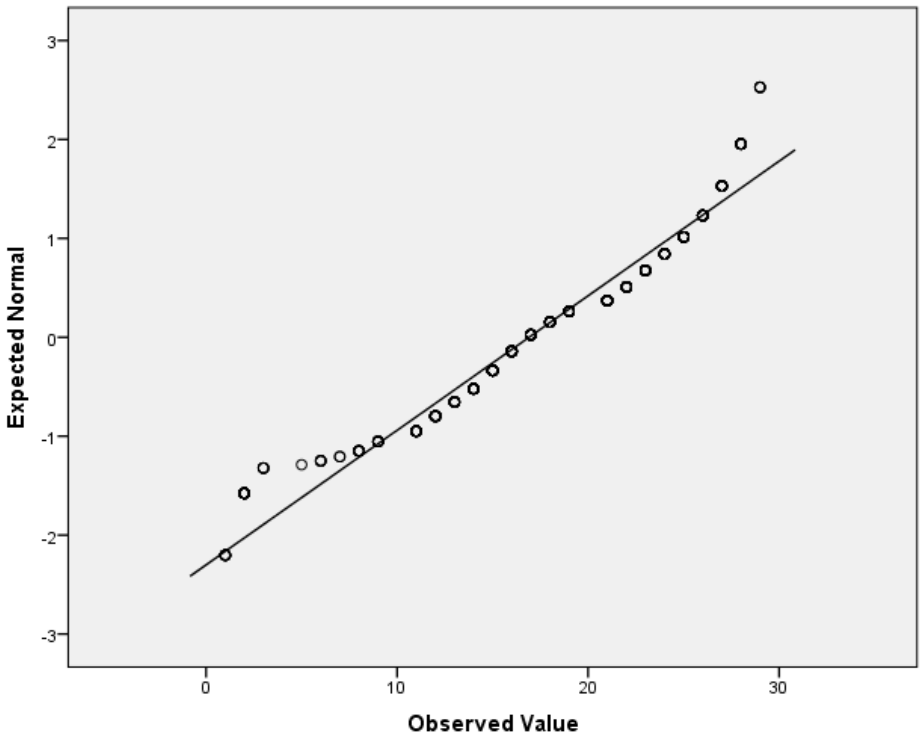

Figure 92: Q-Q Plot of Student Results in Chemistry IB Redeemable Exam 2013

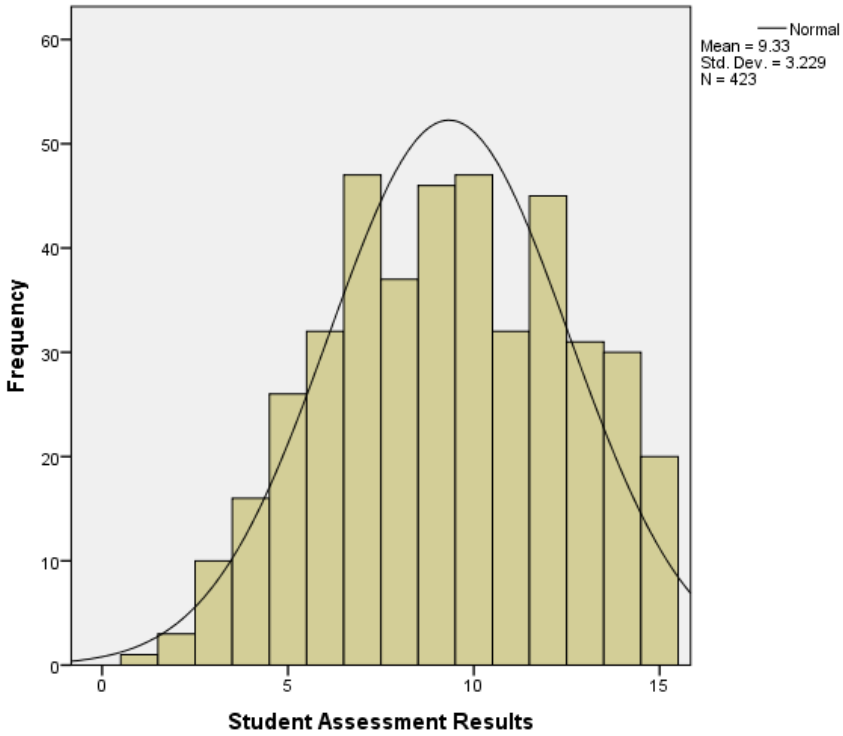

Figure 93: Student Scores Obtained in Chemistry IB Lecture Test 1 2014

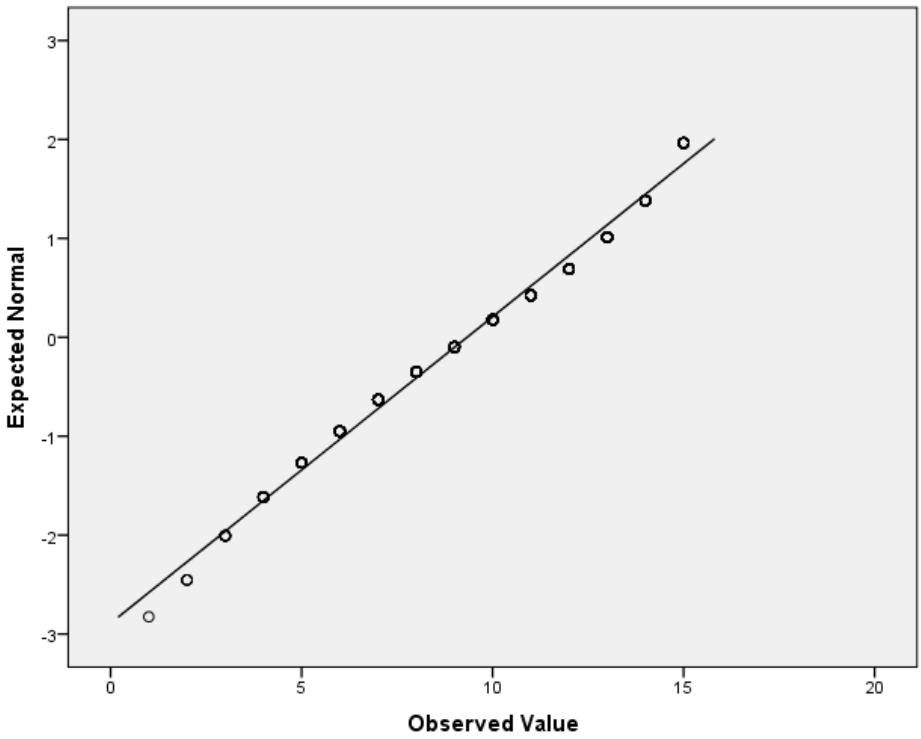

Figure 94: Q-Q Plot of Student Results in Chemistry IB Lecture Test 1 2014

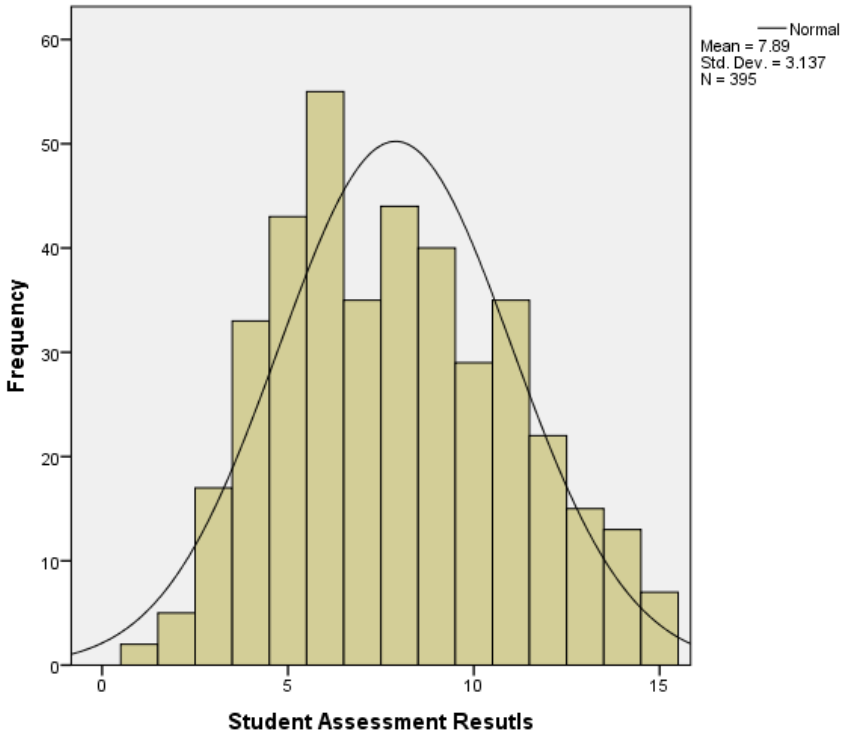

Figure 95: Student Scores Obtained in Chemistry IB Lecture Test 2 2014

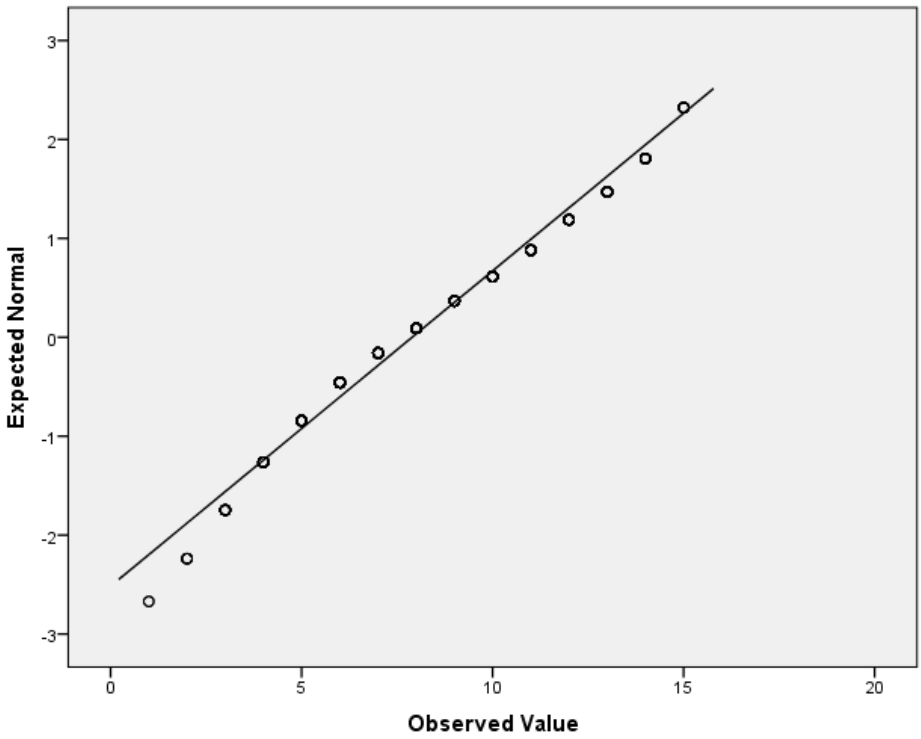

Figure 96: Q-Q Plot of Student Results in Chemistry IB Lecture Test 2 2014

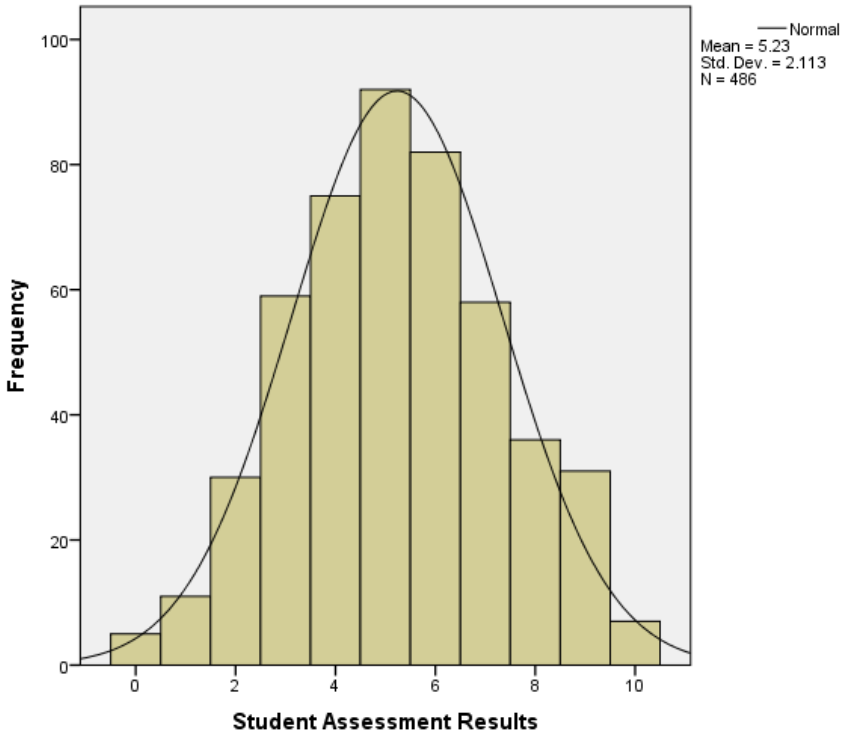

Figure 97: Student Scores Obtained in Chemistry IB Exam 2014

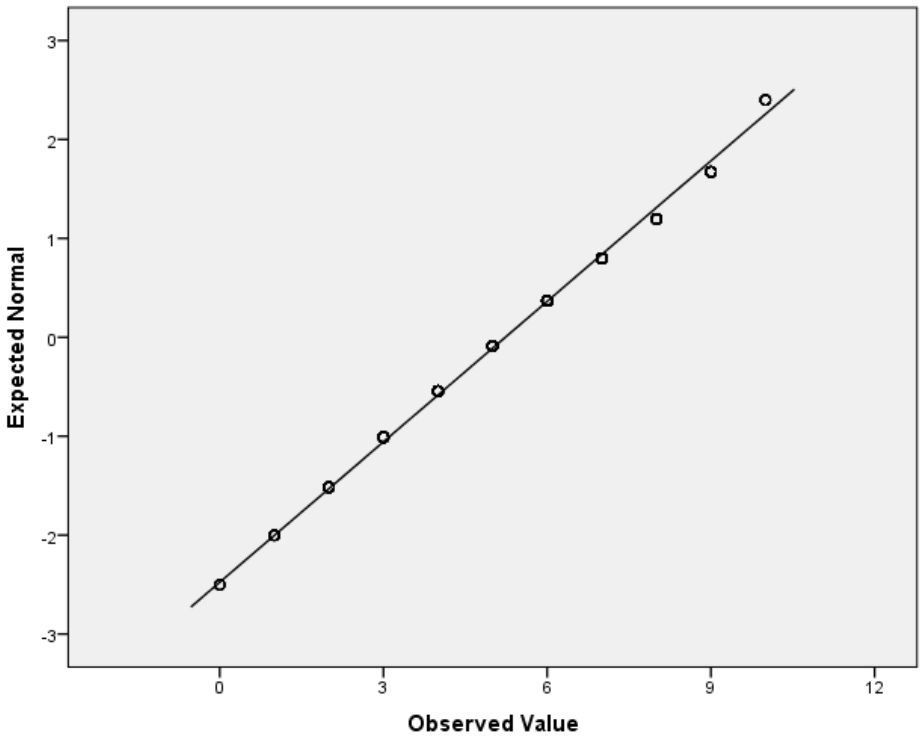

Figure 98: Q-Q Plot of Student Results in Chemistry IB Exam 2014

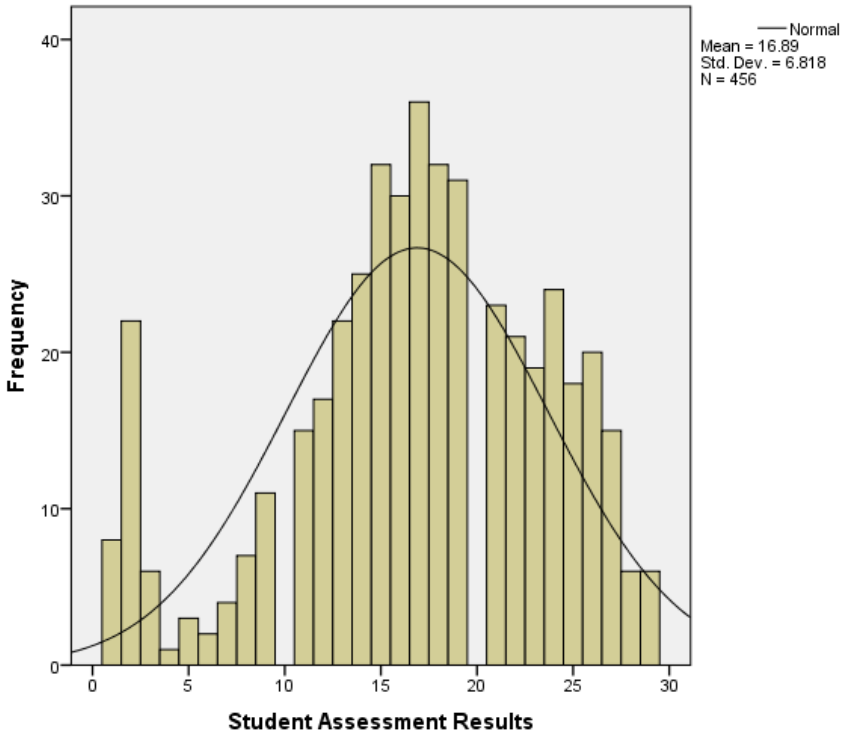

Figure 99: Student Scores Obtained in Chemistry IB Redeemable Exam 2014

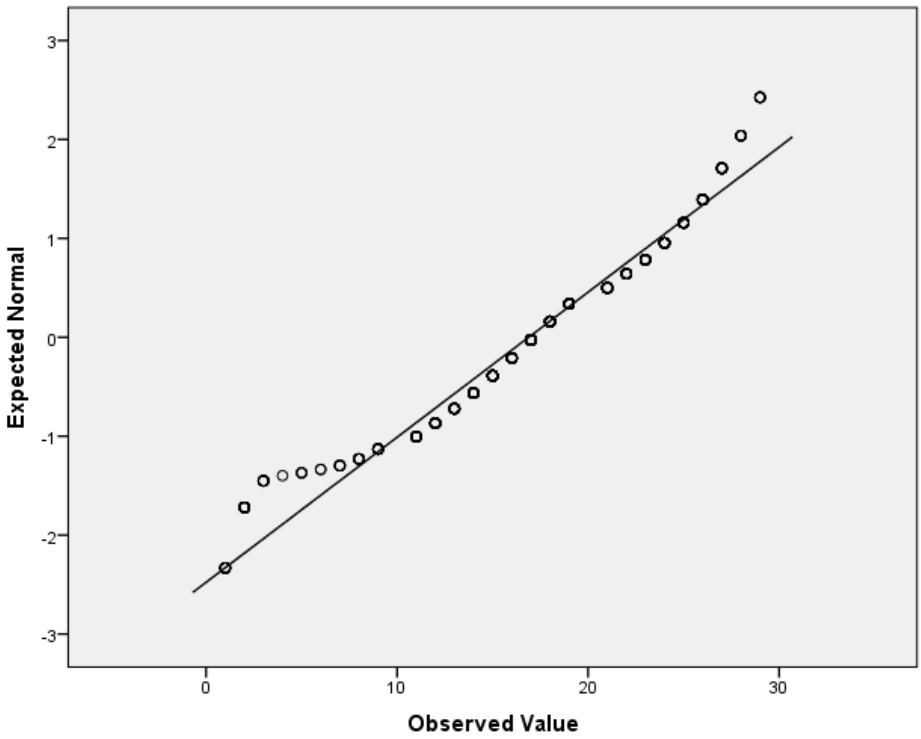

Figure 100: Q-Q Plot of Student Results in Chemistry IB Redeemable Exam 2014

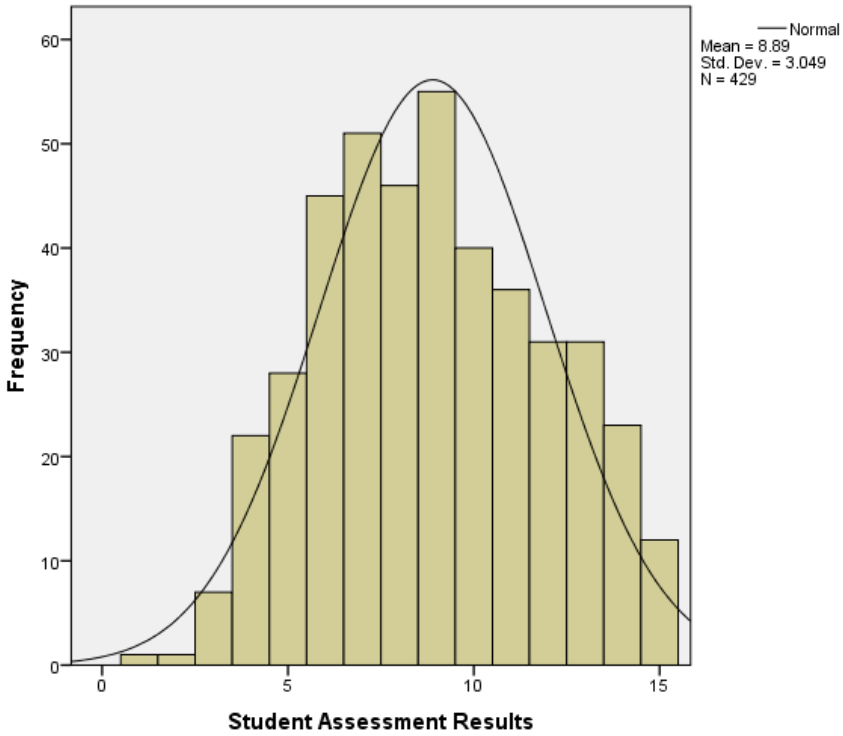

Figure 101: Student Scores Obtained in Chemistry IB Lecture Test 1 2015

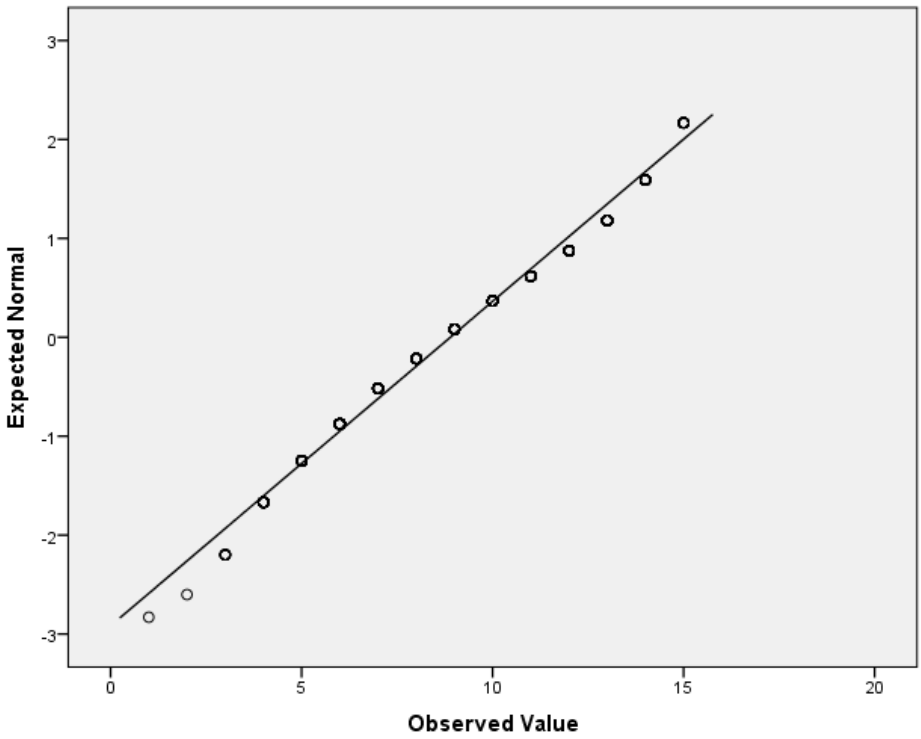

Figure 102: Q-Q Plot of Student Results in Chemistry IB Lecture Test 1 2015

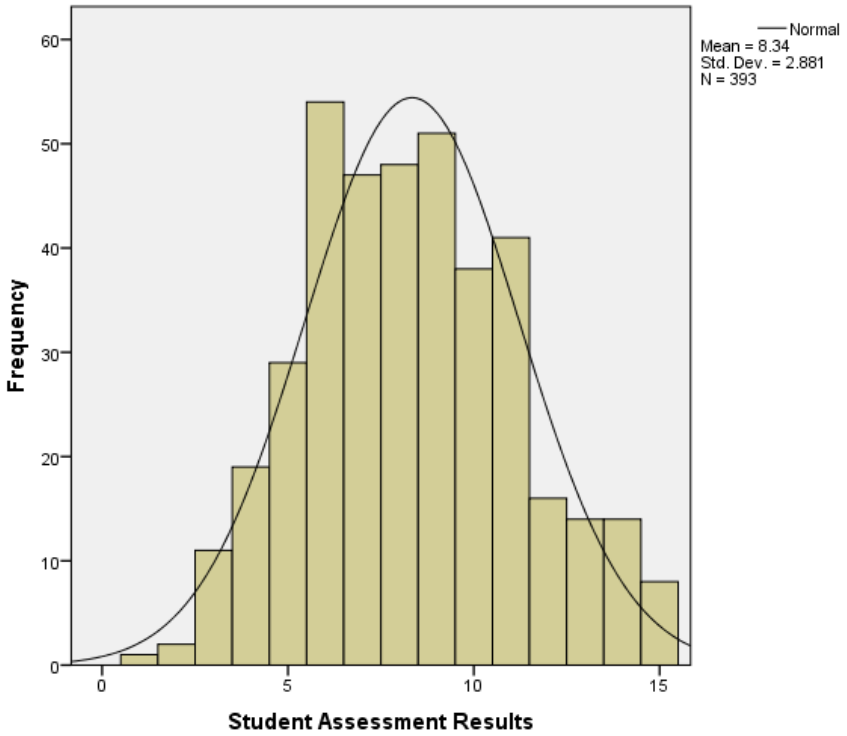

Figure 103: Student Scores Obtained in Chemistry IB Lecture Test 2 2015

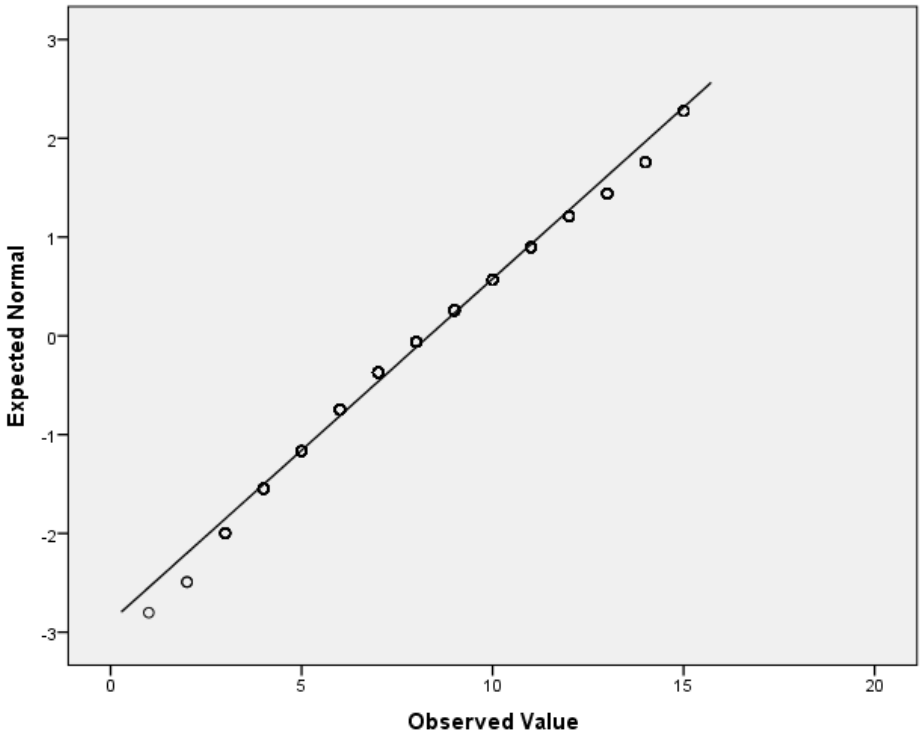

Figure 104: Q-Q Plot of Student Results in Chemistry IB Lecture Test 2 2015

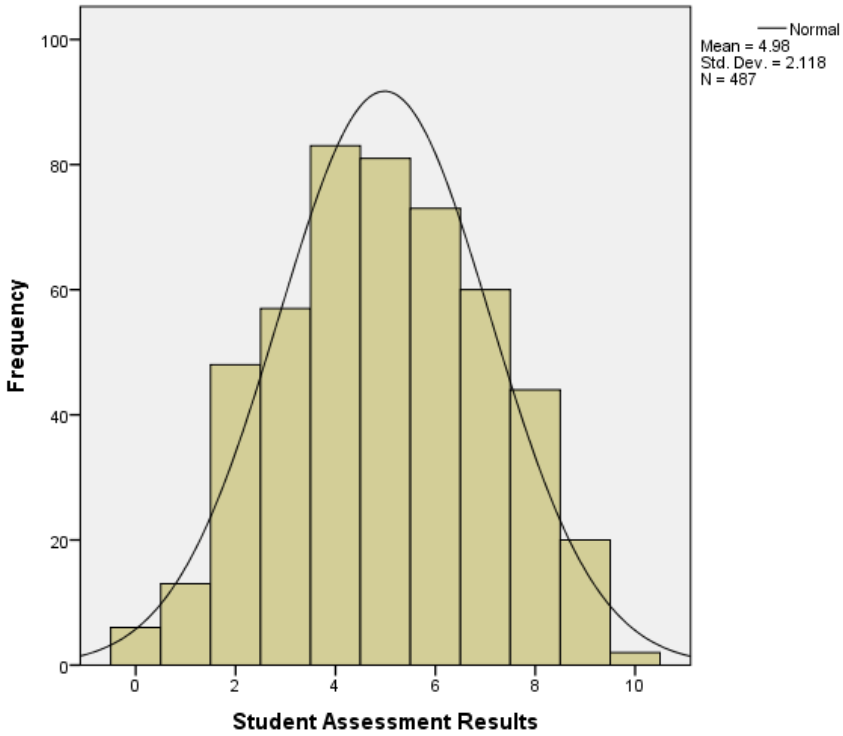

Figure 105: Student Scores Obtained in Chemistry IB Exam 2015

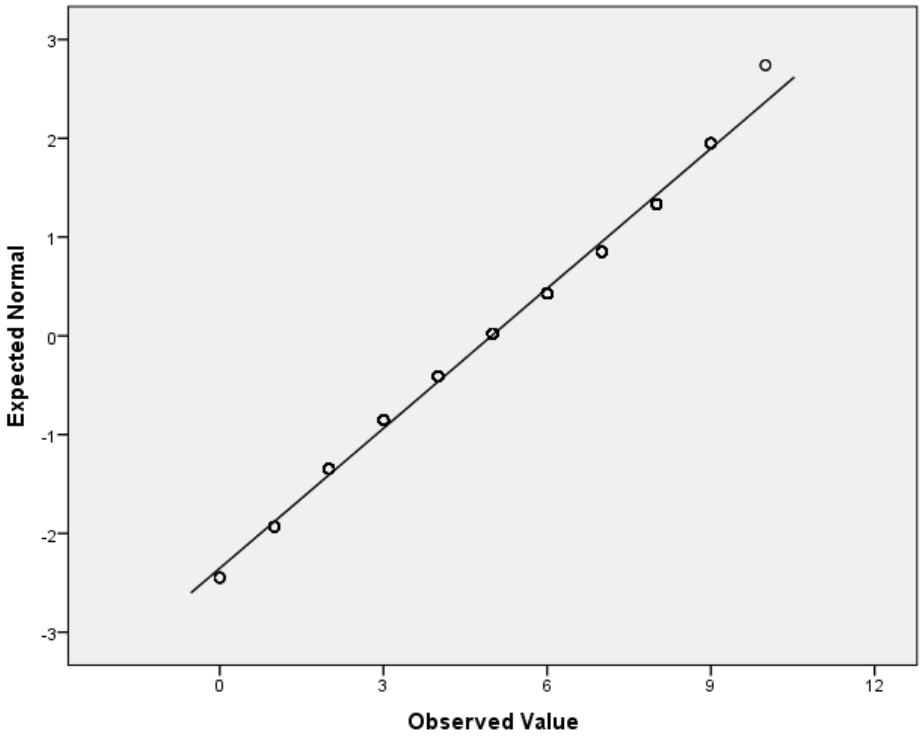

Figure 106: Q-Q Plot of Student Results in Chemistry IB Exam 2015

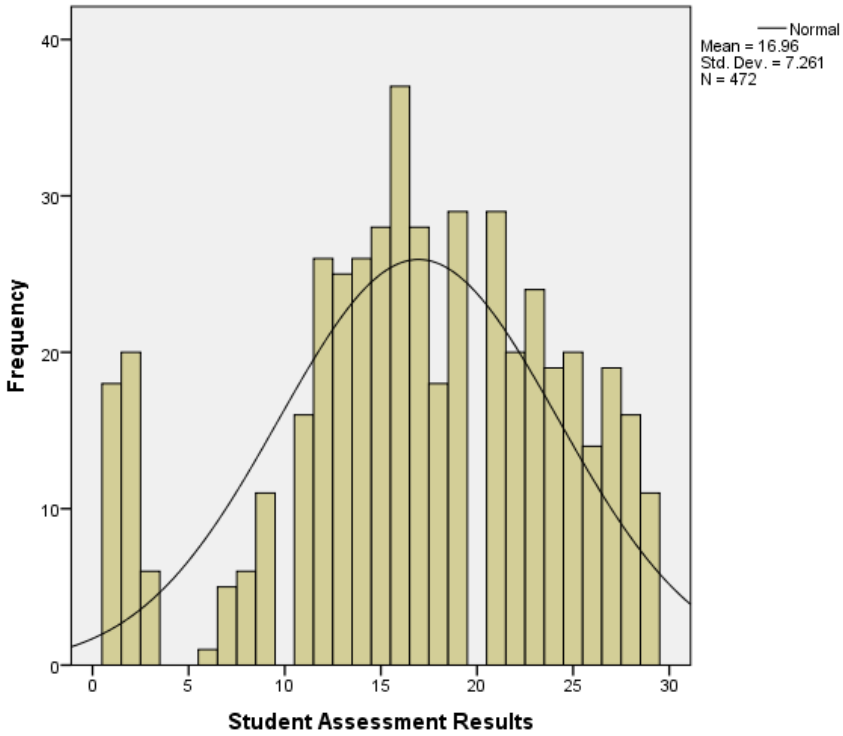

Figure 107: Student Scores Obtained in Chemistry IB Redeemable Exam 2015

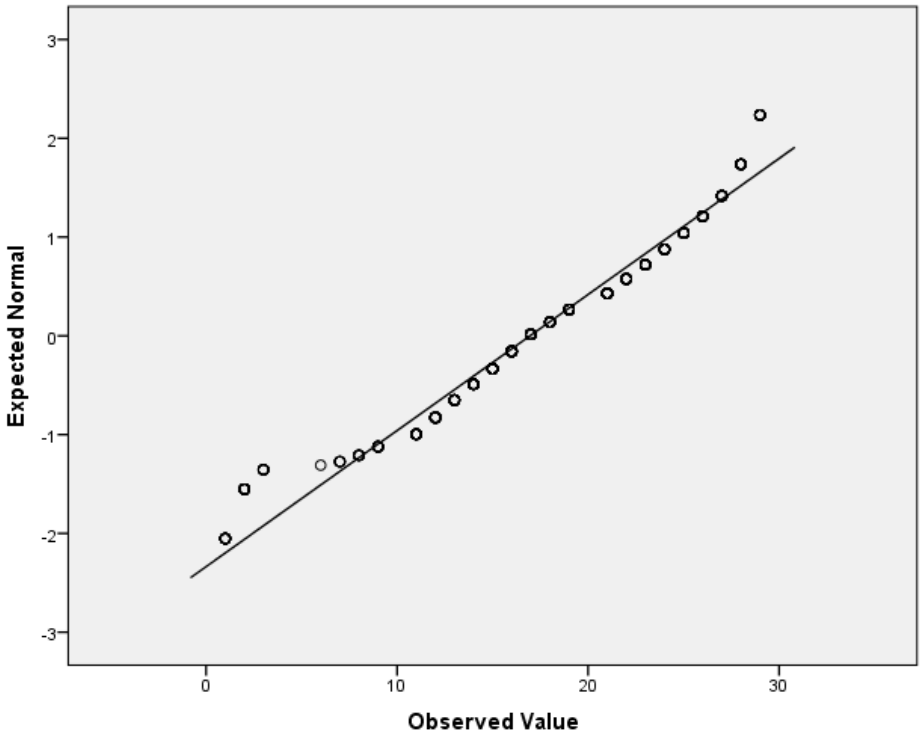

Figure 108: Q-Q Plot of Student Results in Chemistry IB Redeemable Exam 2015

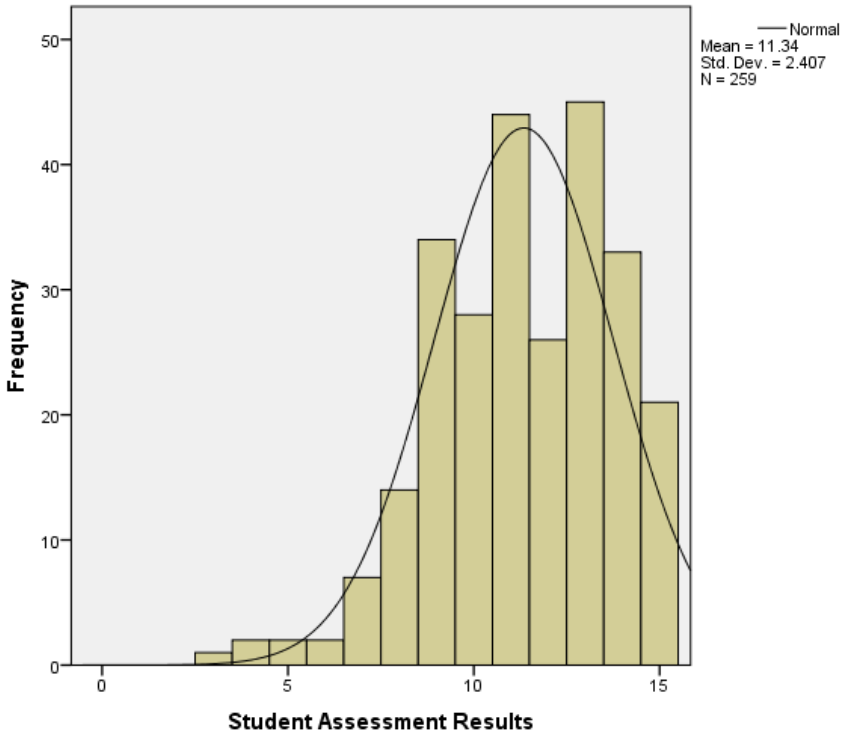

Figure 109: Student Scores Obtained in Foundations of Chemistry IA Lecture Test 1 2012

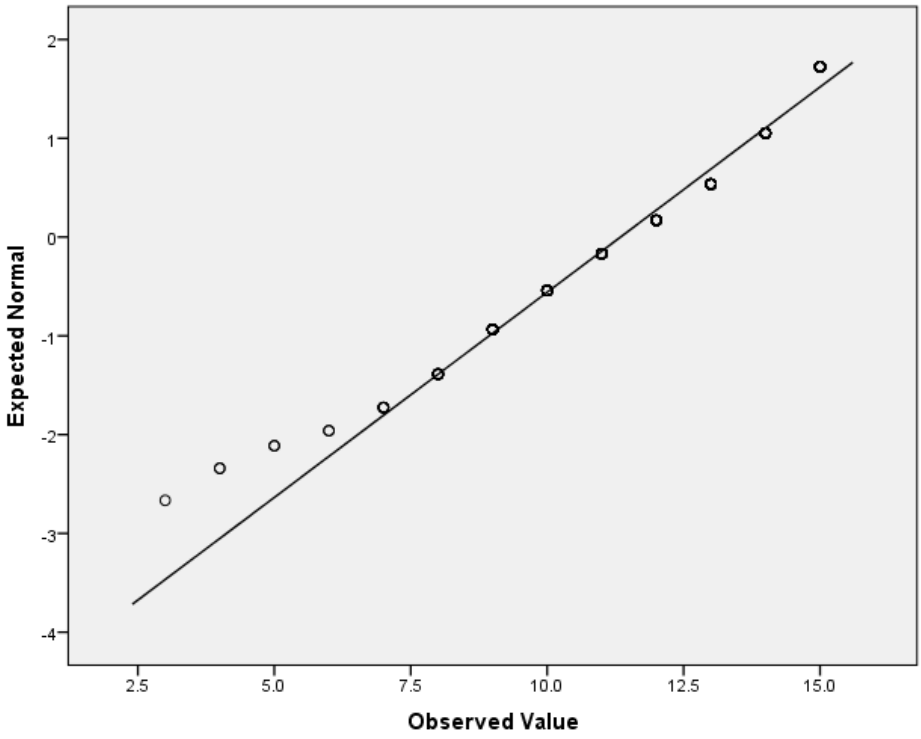

Figure 110: Q-Q Plot of Student Results in Foundations of Chemistry IA Lecture Test 1 2012

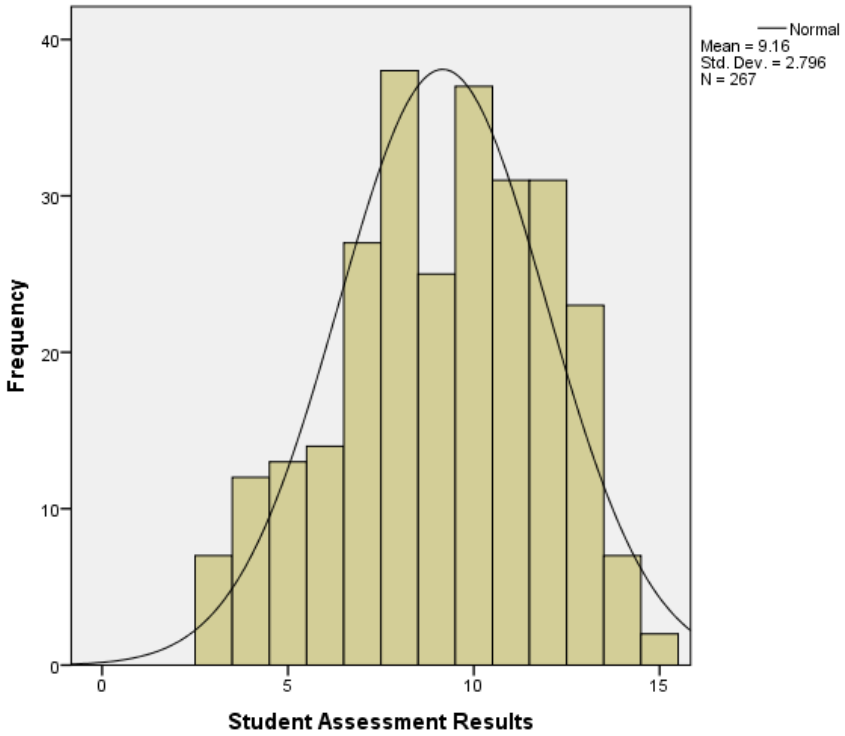

Figure 111: Student Scores Obtained in Foundations of Chemistry IA Lecture Test 2 2012

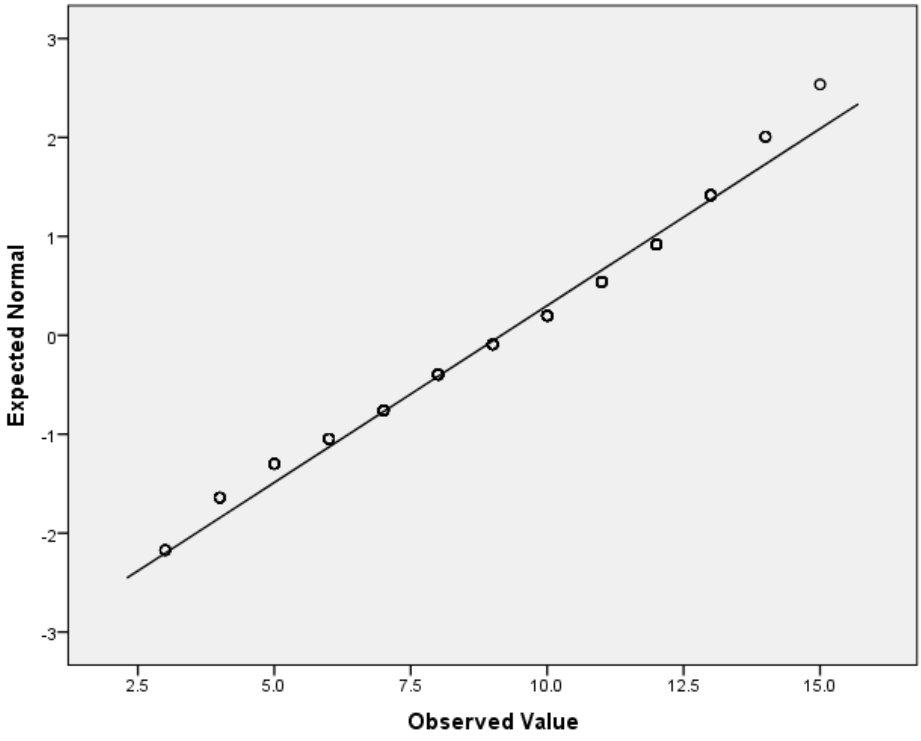

Figure 112: Q-Q Plot of Student Results in Foundations of Chemistry IA Lecture Test 2 2012

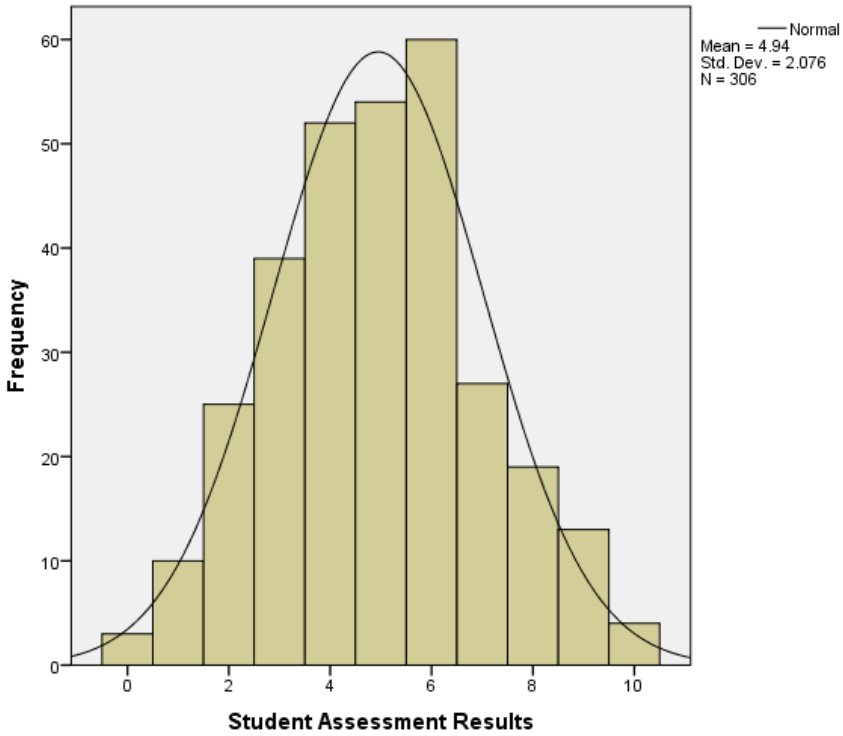

Figure 113: Student Scores Obtained in Foundations of Chemistry IA Exam 2012

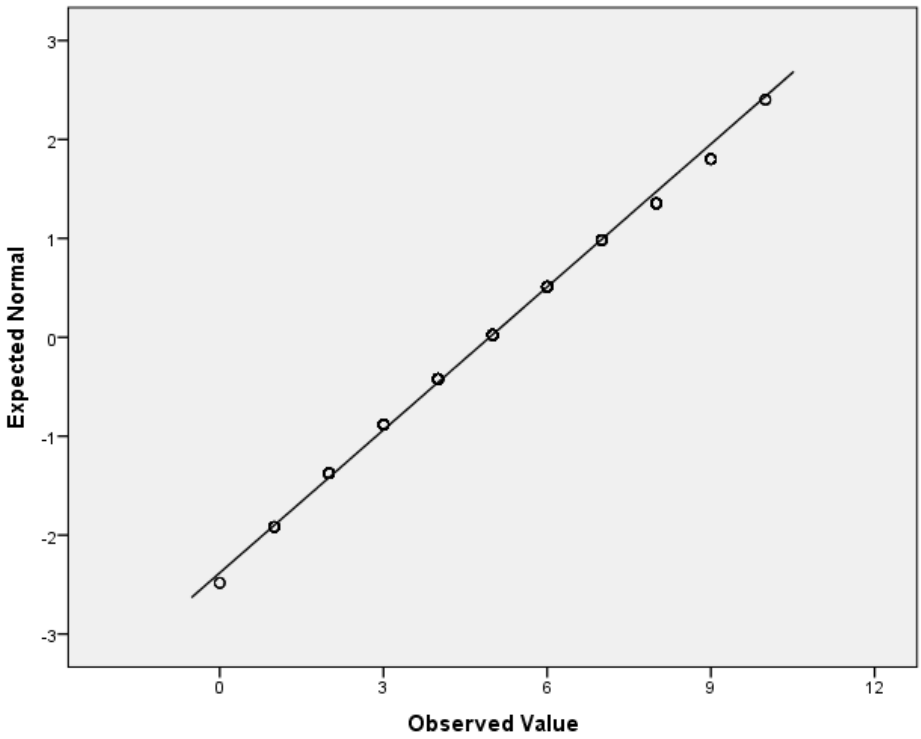

Figure 114: Q-Q Plot of Student Results in Foundations of Chemistry IA Exam 2012

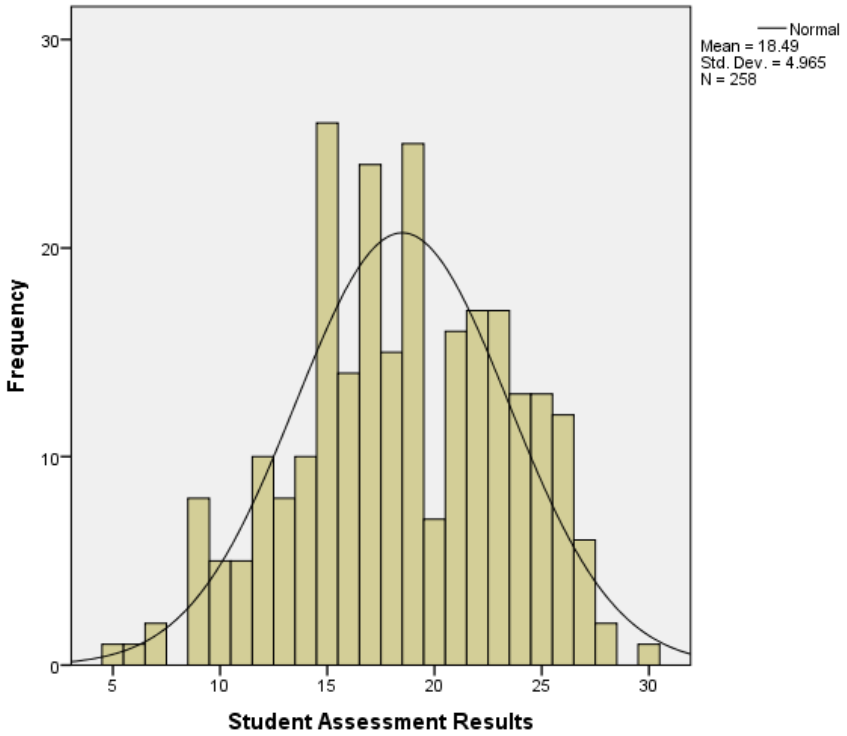

Figure 115: Student Scores Obtained in Foundations of Chemistry IA Redeemable Exam 2012

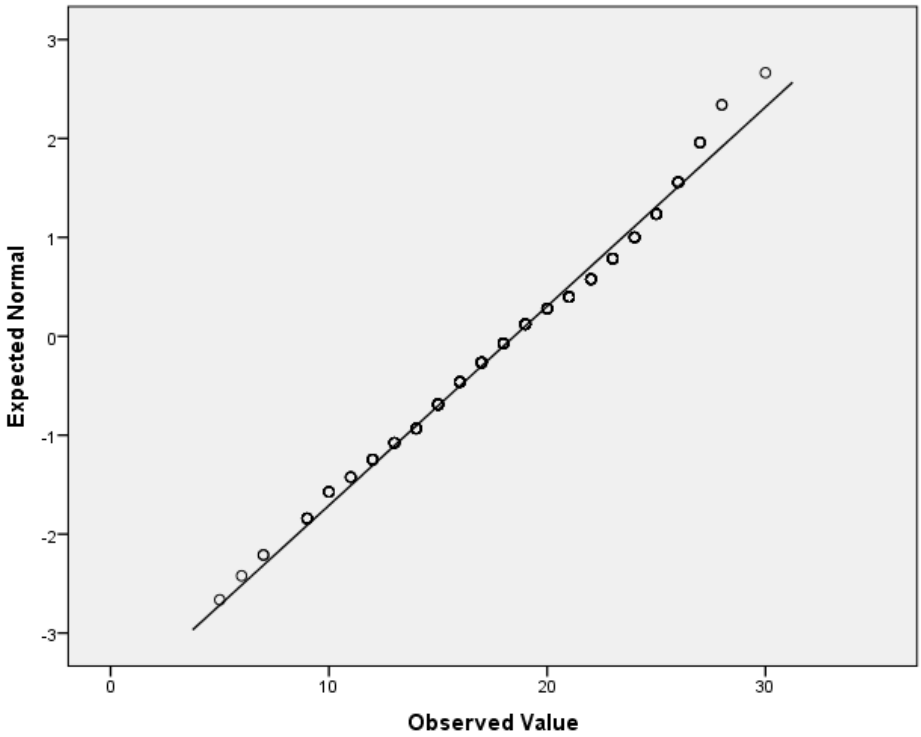

Figure 116: Q-Q Plot of Student Results in Foundations of Chemistry IA Redeemable Exam 2012

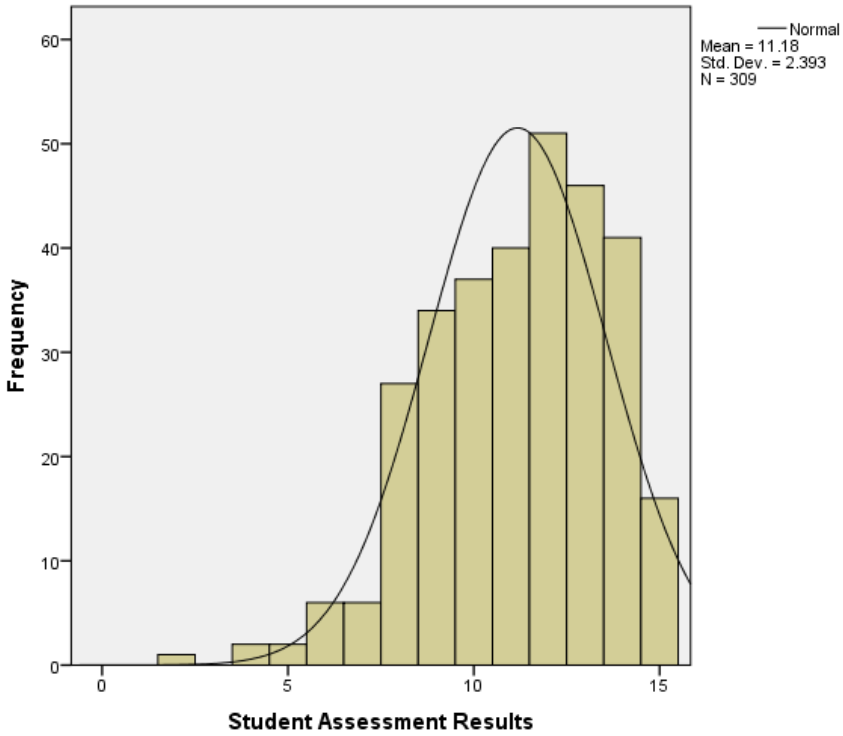

Figure 117: Student Scores Obtained in Foundations of Chemistry IA Lecture Test 1 2013

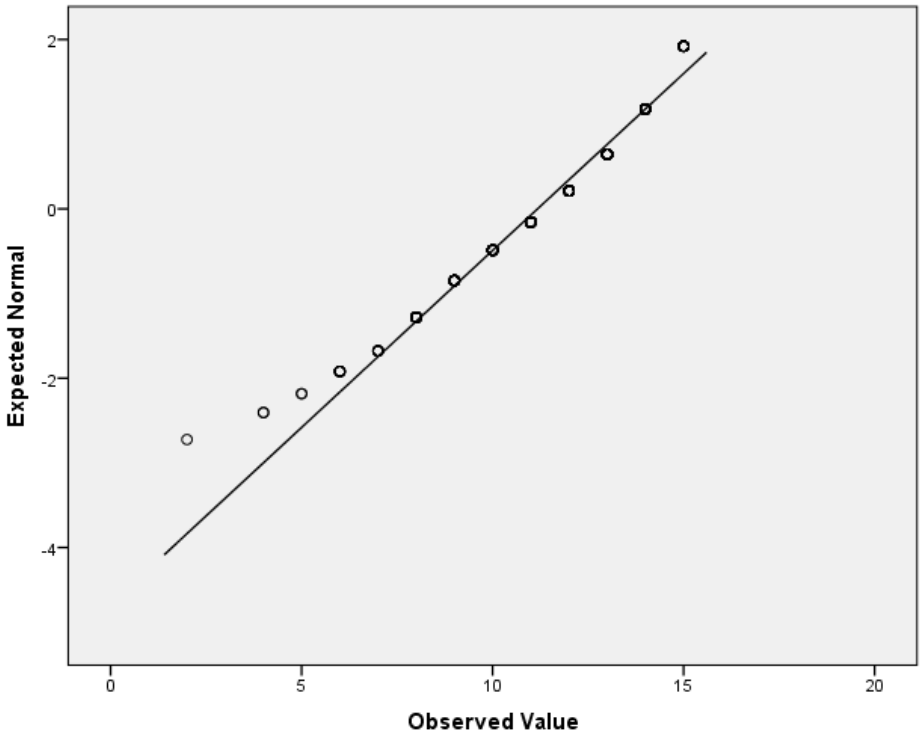

Figure 118: Q-Q Plot of Student Results in Foundations of Chemistry IA Lecture Test 1 2013

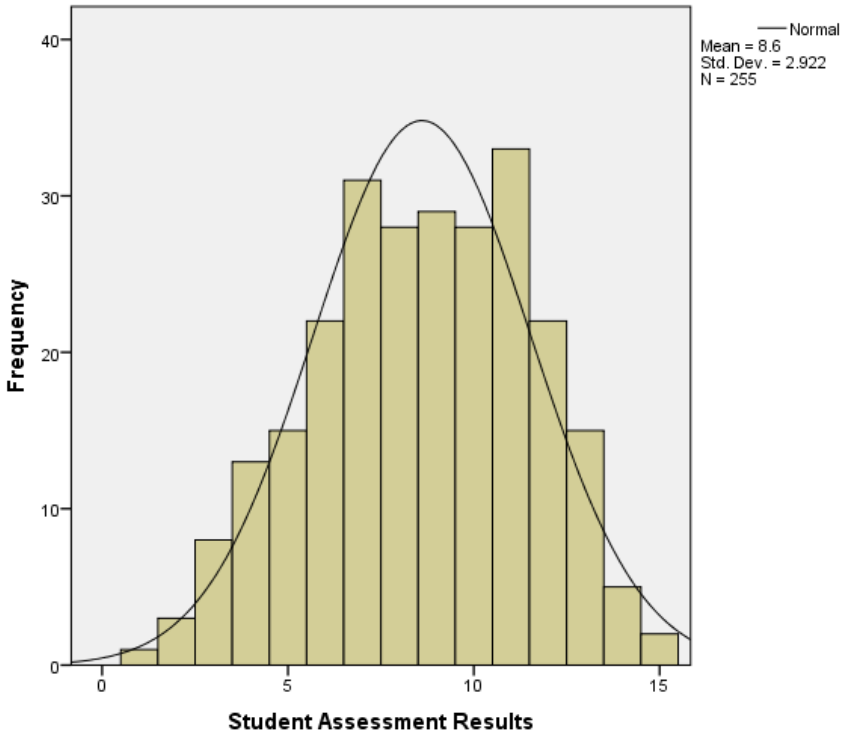

Figure 119: Student Scores Obtained in Foundations of Chemistry IA Lecture Test 2 2013

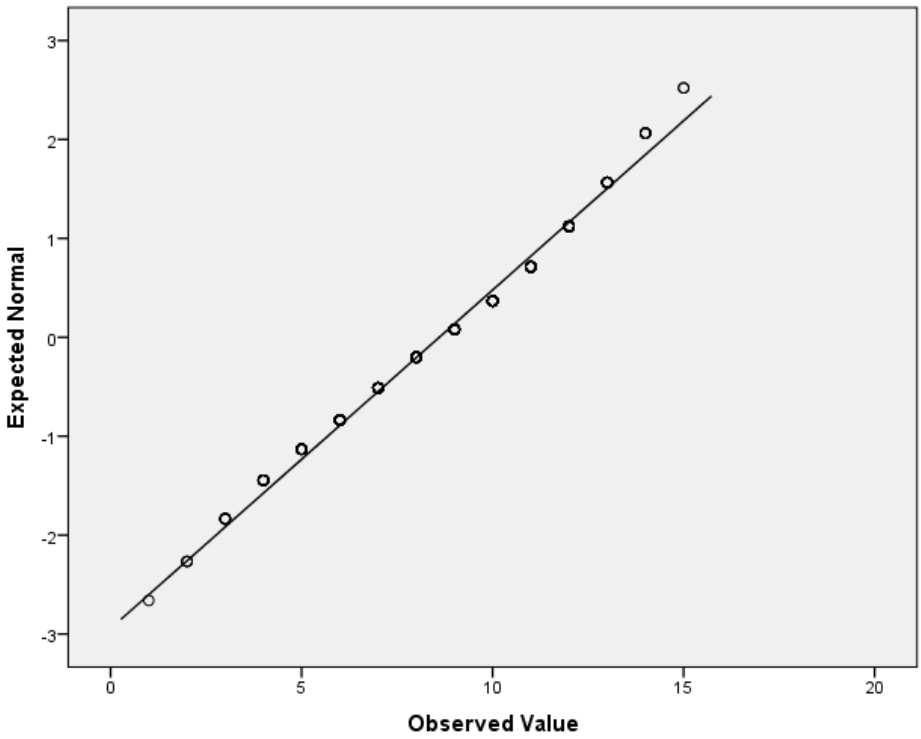

Figure 120: Q-Q Plot of Student Results in Foundations of Chemistry IA Lecture Test 2 2013

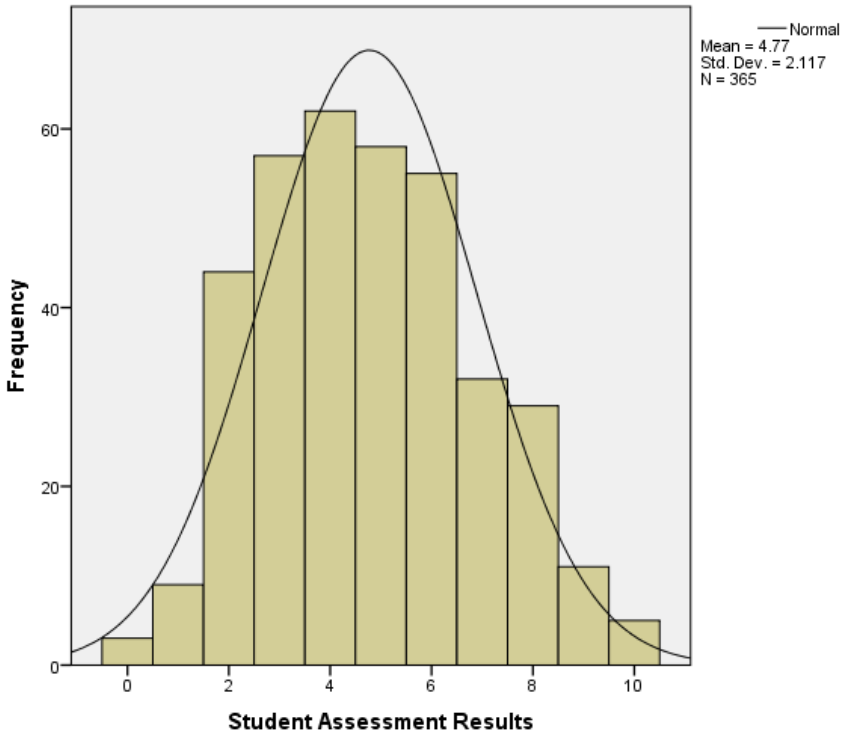

Figure 121: Student Scores Obtained in Foundations of Chemistry IA Exam 2013

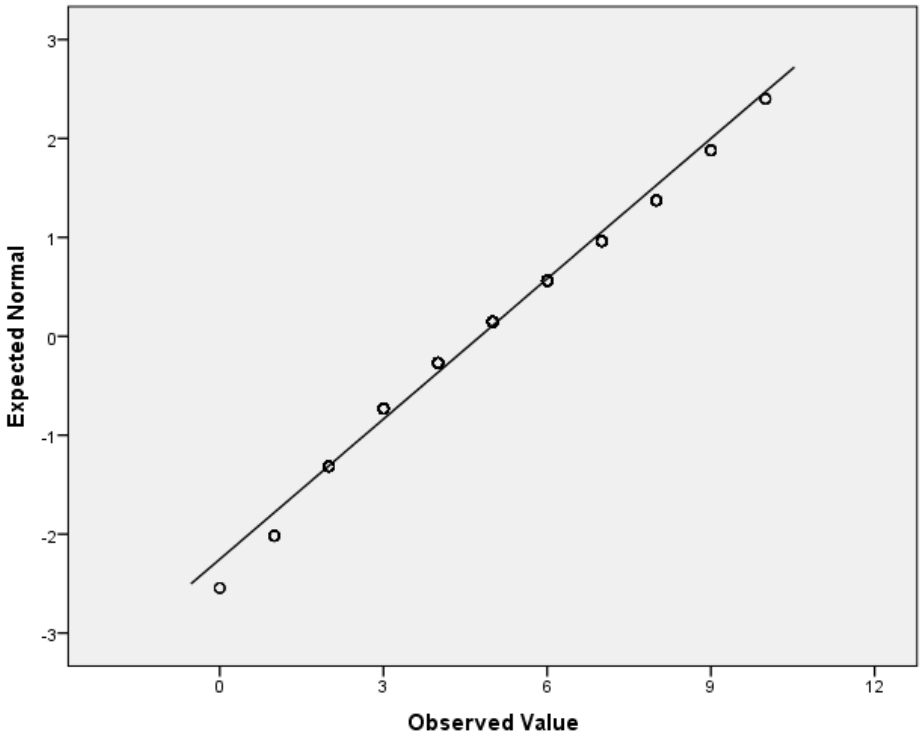

Figure 122: Q-Q Plot of Student Results in Foundations of Chemistry IA Exam 2013

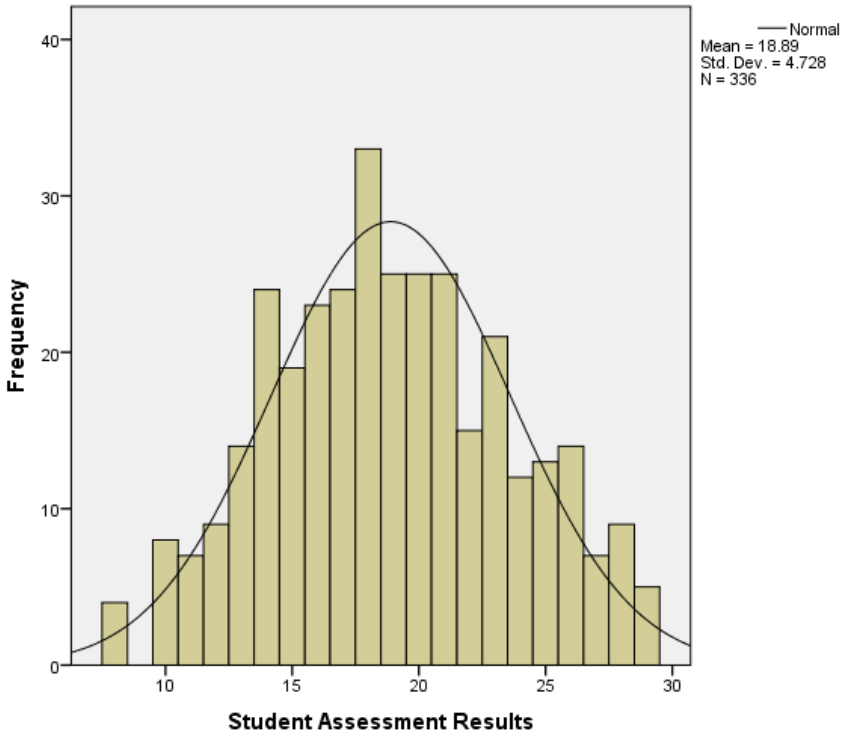

Figure 123: Student Scores Obtained in Foundations of Chemistry IA Redeemable Exam 2013

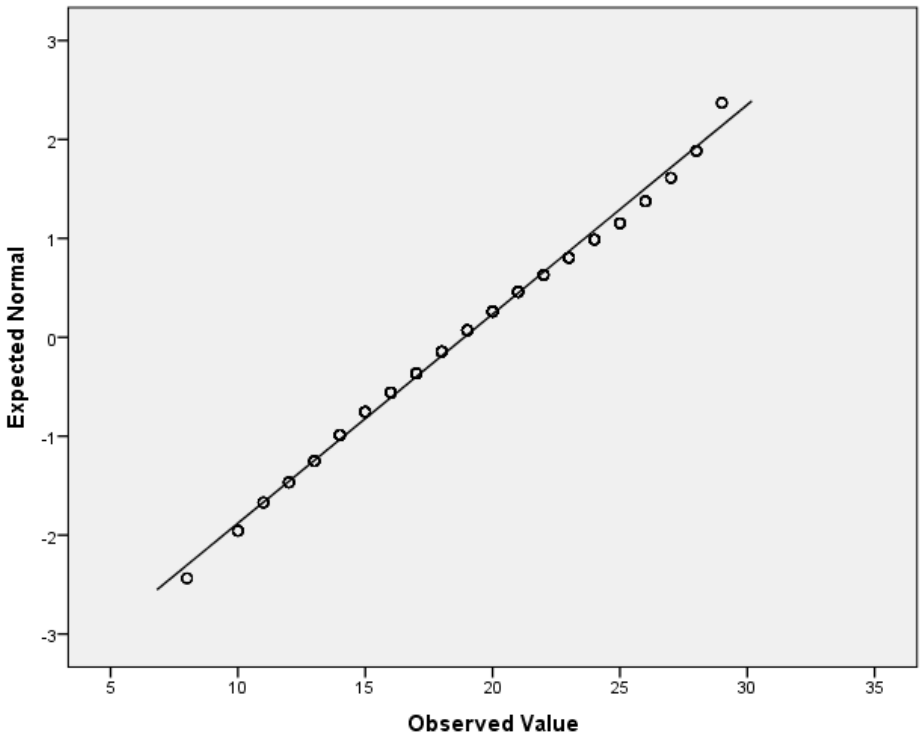

Figure 124: Q-Q Plot of Student Results in Foundations of Chemistry IA Redeemable Exam 2013

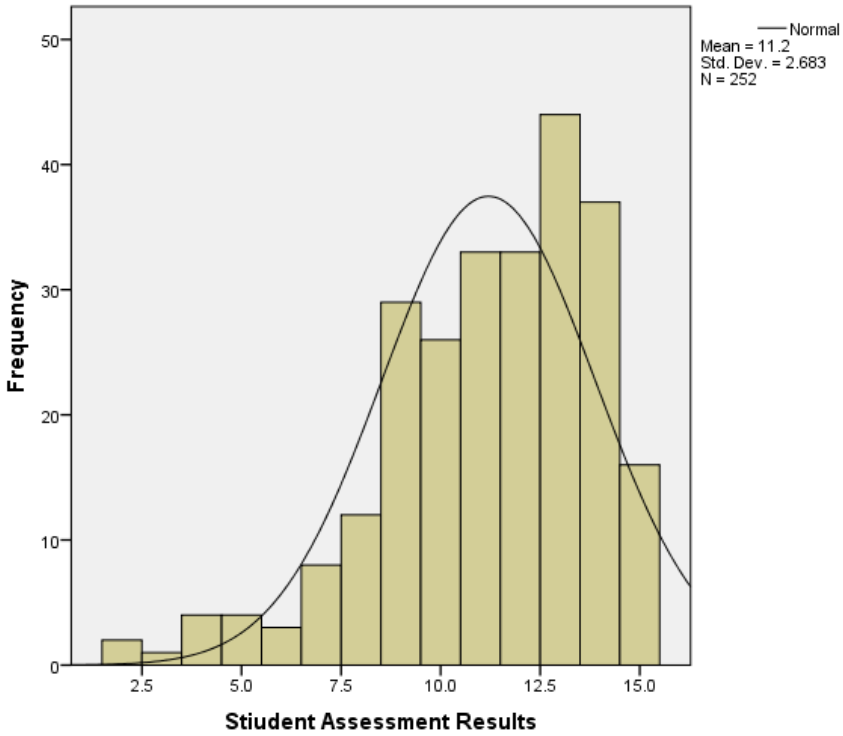

Figure 125: Student Scores Obtained in Foundations of Chemistry IA Lecture Test 1 2014

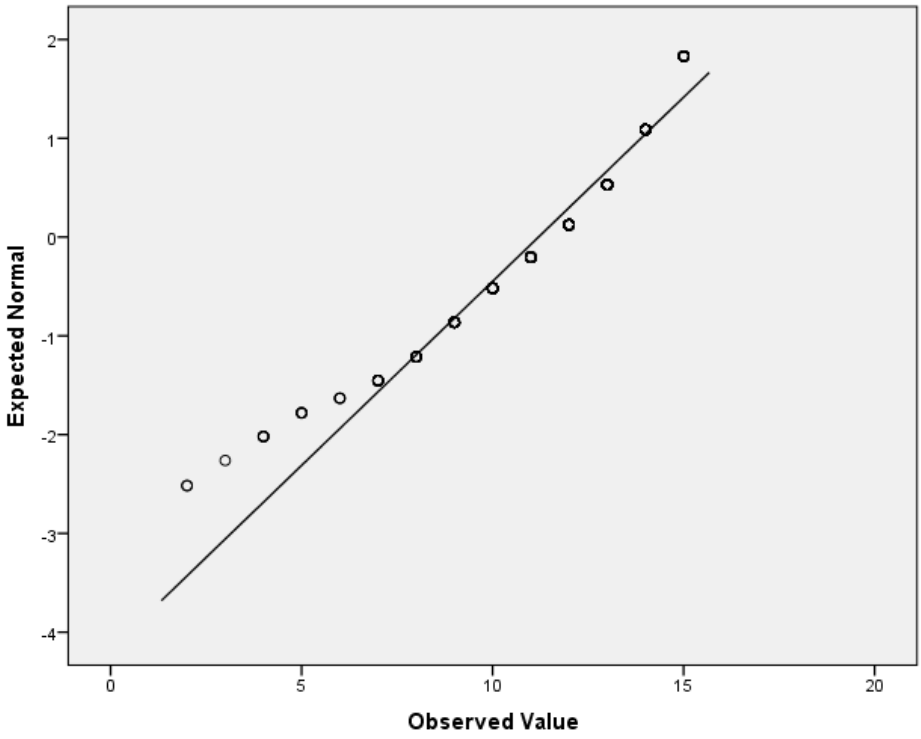

Figure 126: Q-Q Plot of Student Results in Foundations of Chemistry IA Lecture Test 1 2014

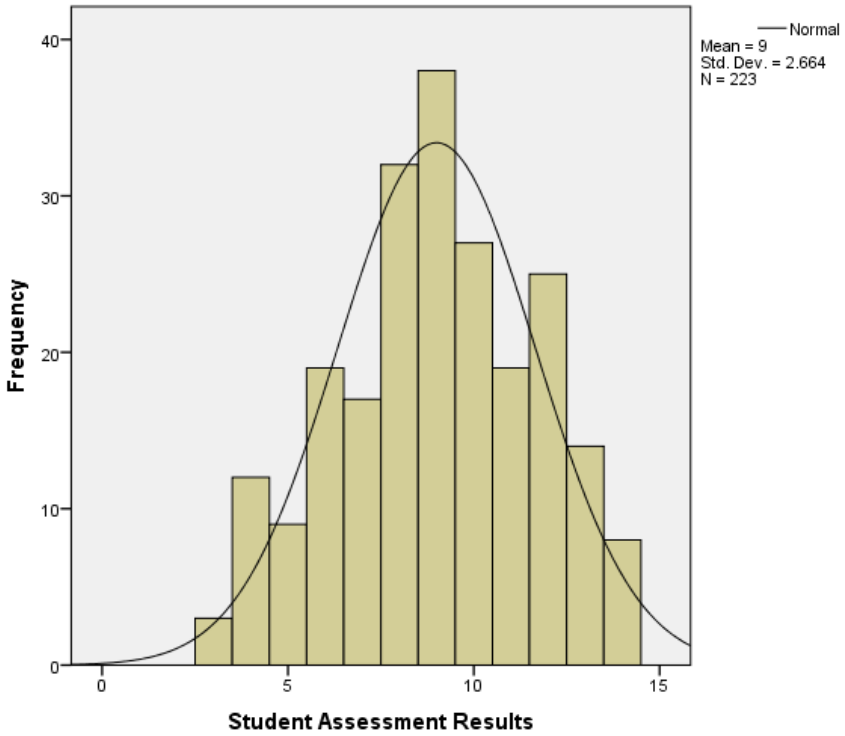

Figure 127: Student Scores Obtained in Foundations of Chemistry IA Lecture Test 2 2014

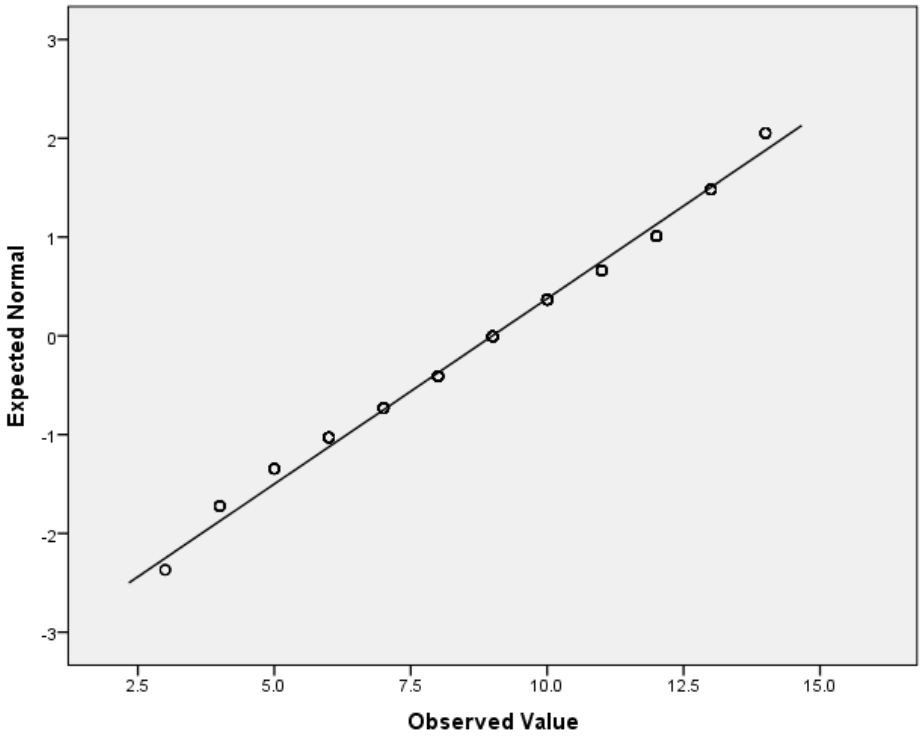

Figure 128: Q-Q Plot of Student Results in Foundations of Chemistry IA Lecture Test 2 2014

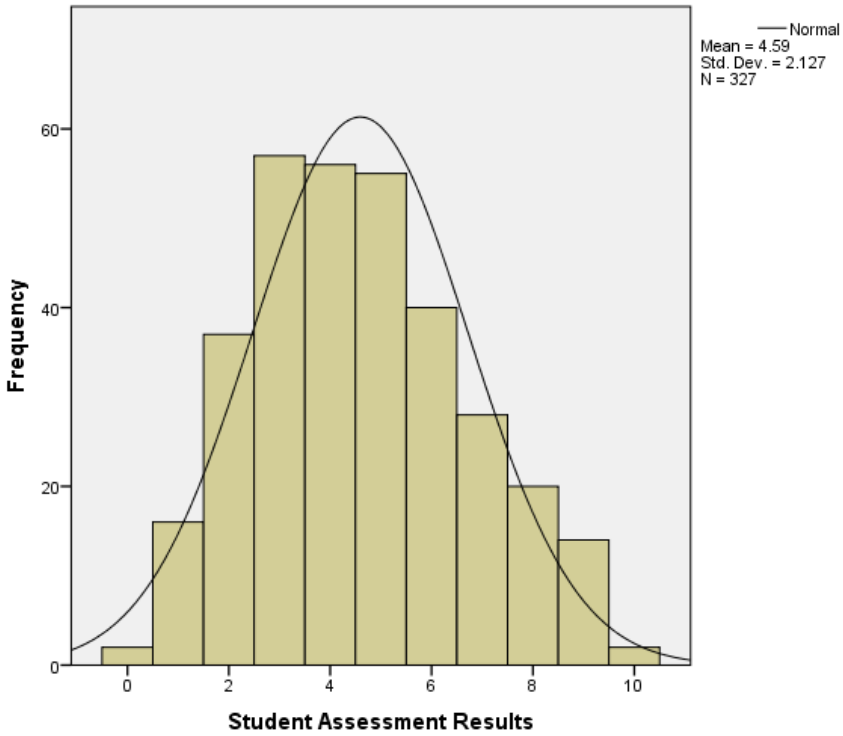

Figure 129: Student Scores Obtained in Foundations of Chemistry IA Exam 2014

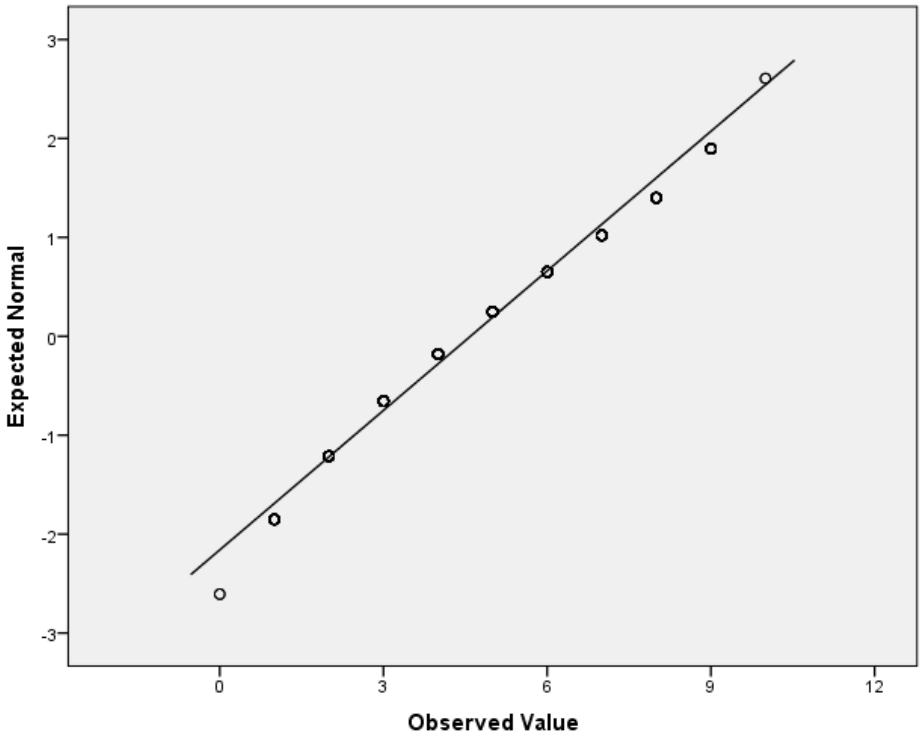

Figure 130: Q-Q Plot of Student Results in Foundations of Chemistry IA Exam 2014

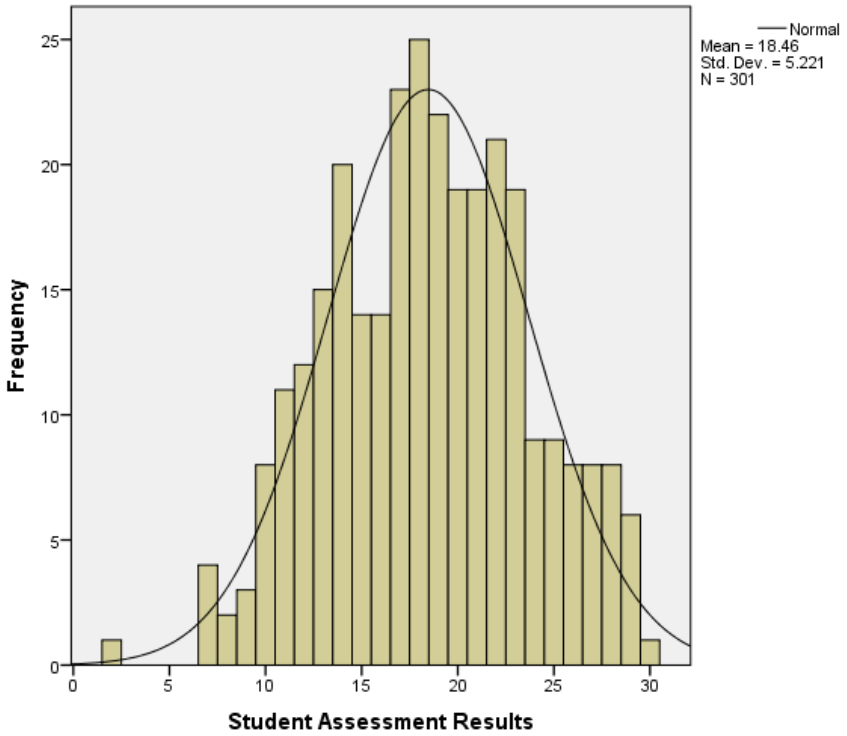

Figure 131: Student Scores Obtained in Foundations of Chemistry IA Redeemable Exam 2014

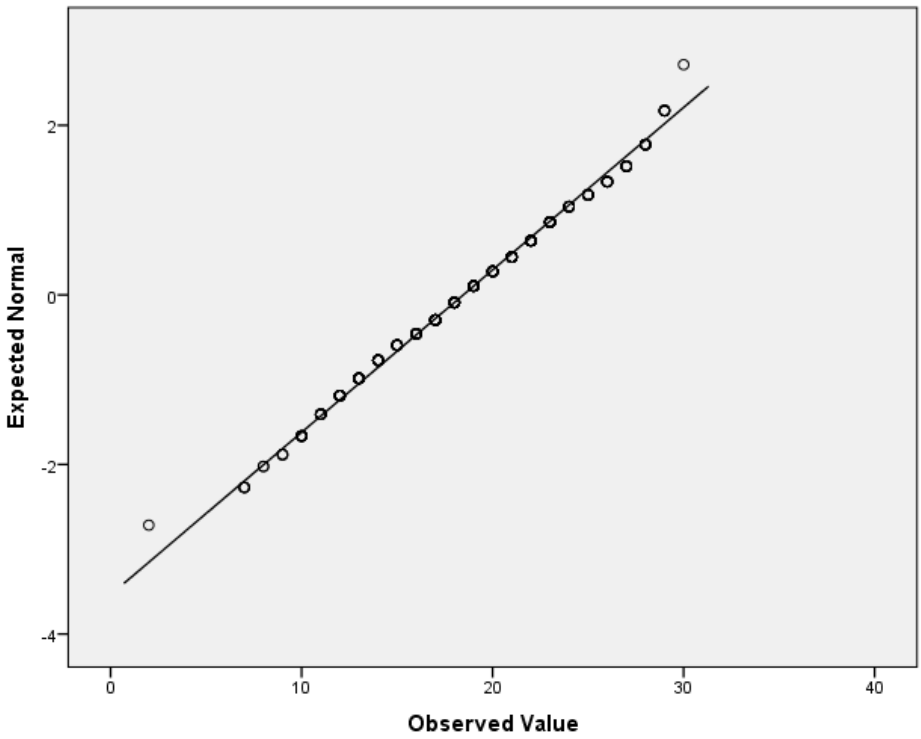

Figure 132: Q-Q Plot of Student Results in Foundations of Chemistry IA Redeemable Exam 2014

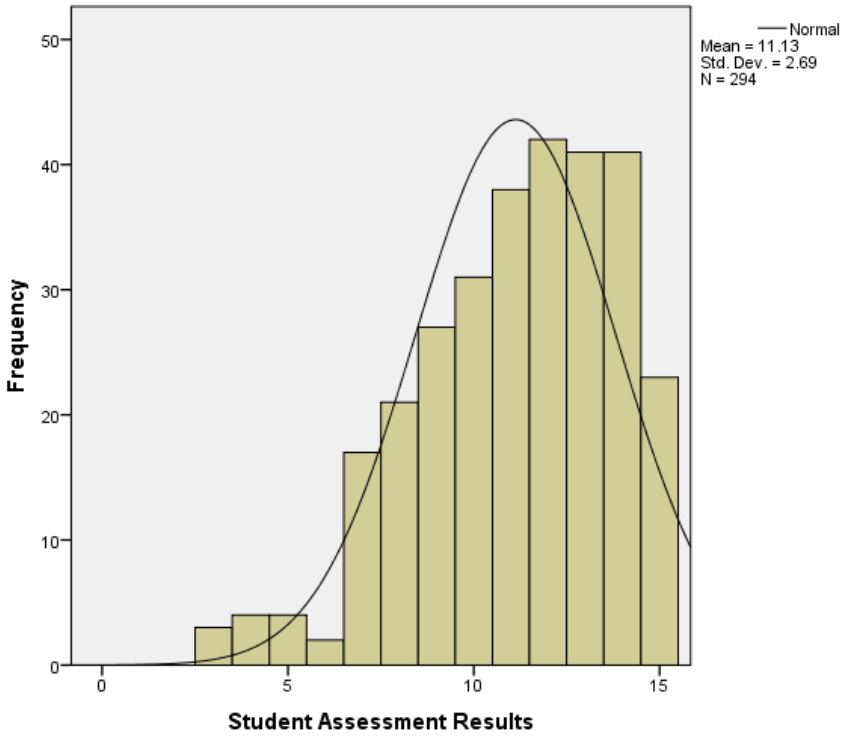

Figure 133: Student Scores Obtained in Foundations of Chemistry IA Lecture Test 1 2015

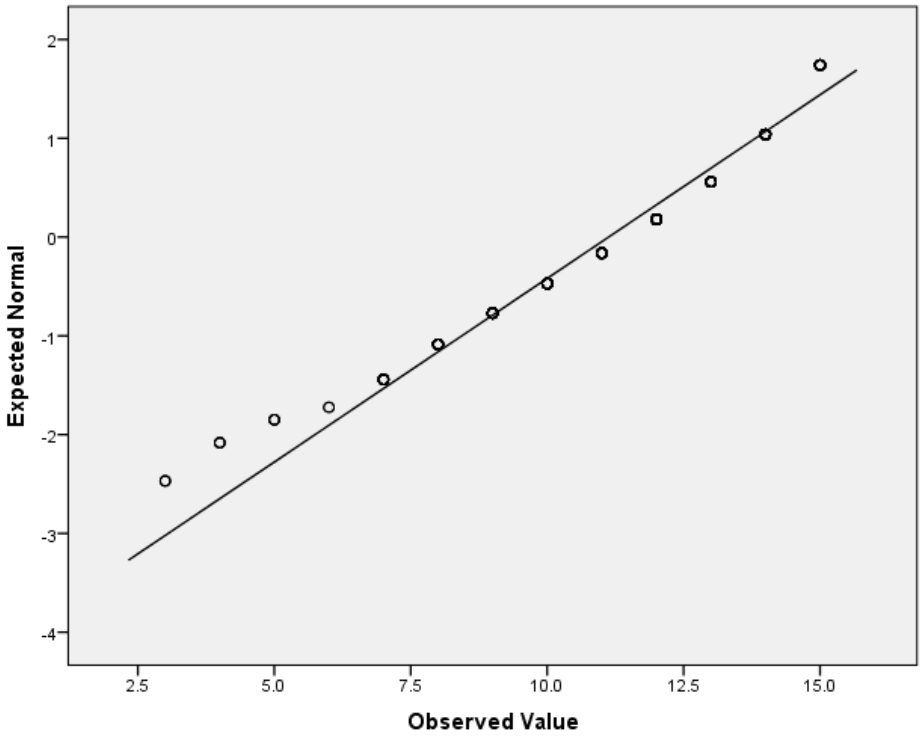

Figure 134: Foundations of Chemistry IA Lecture Test 1 2015

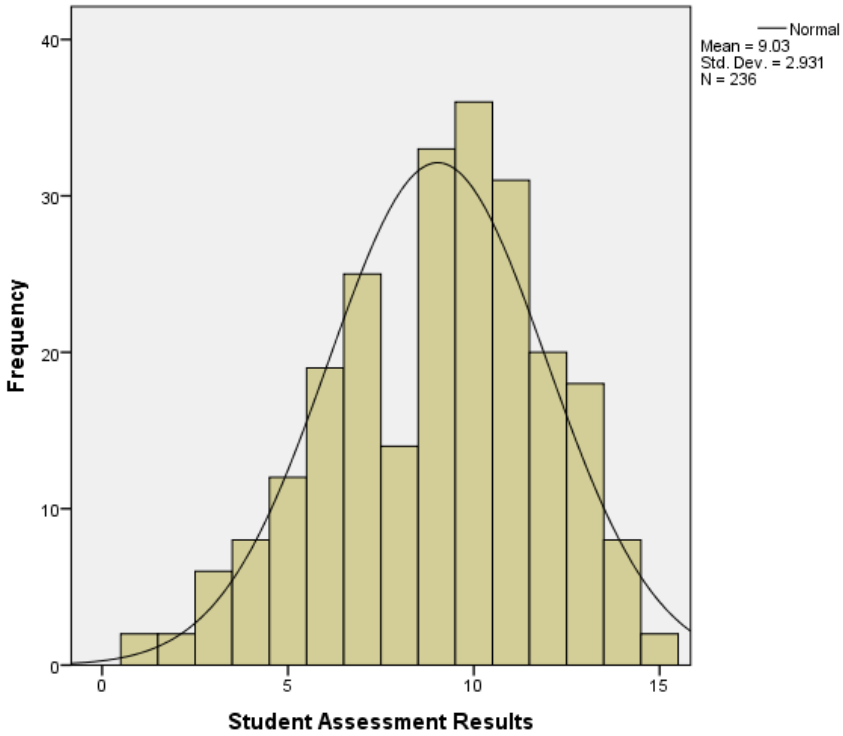

Figure 135: Student Scores Obtained in Foundations of Chemistry IA Lecture Test 2 2015

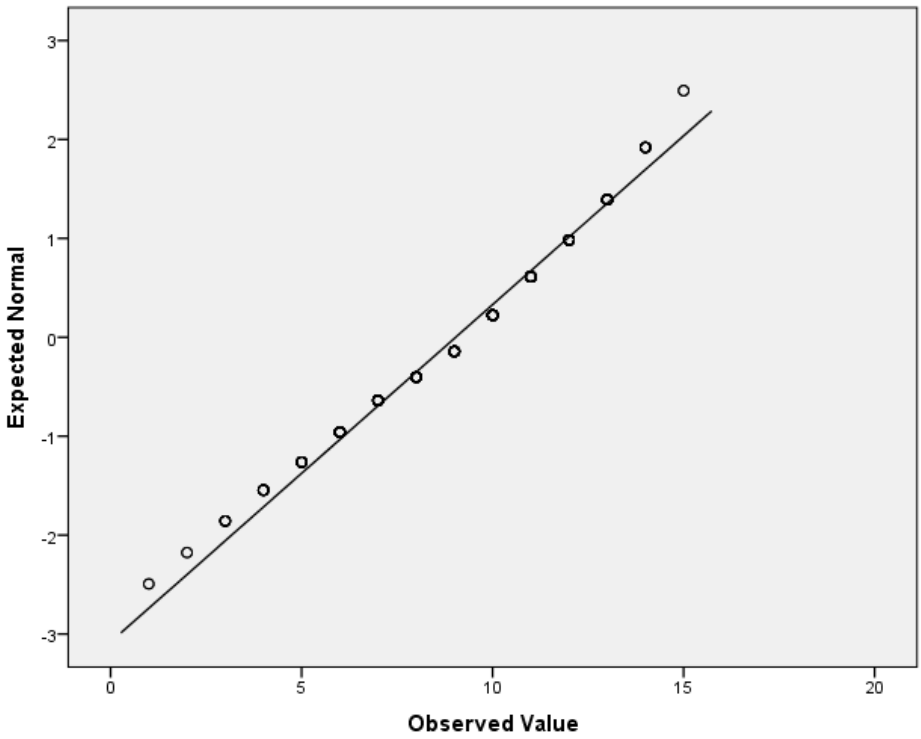

Figure 136: Q-Q Plot of Student Results in Foundations of Chemistry IA Lecture Test 2 2015

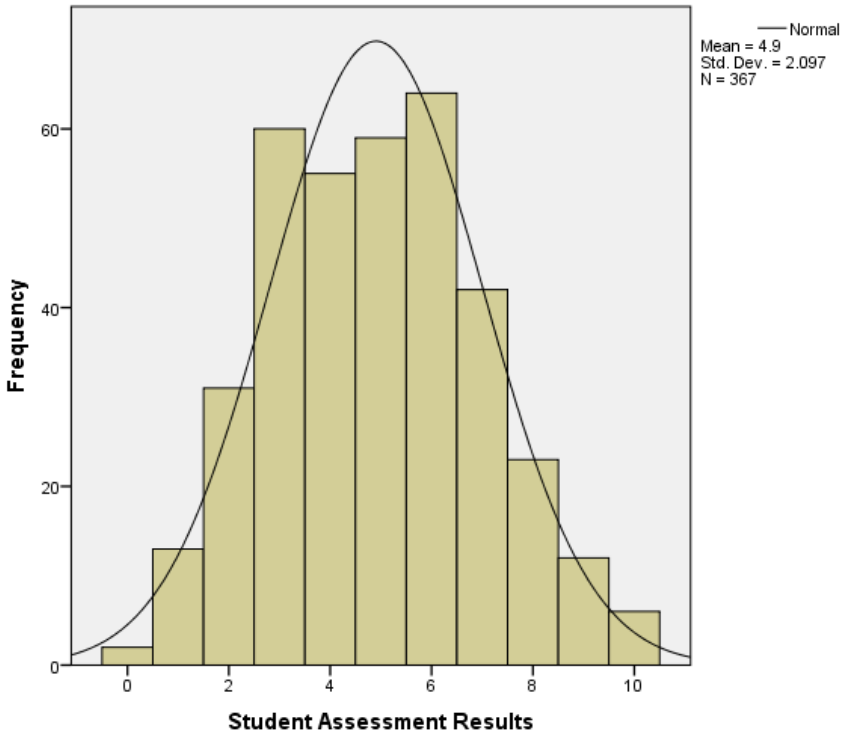

Figure 137: Student Scores Obtained in Foundations of Chemistry IA Exam 2015

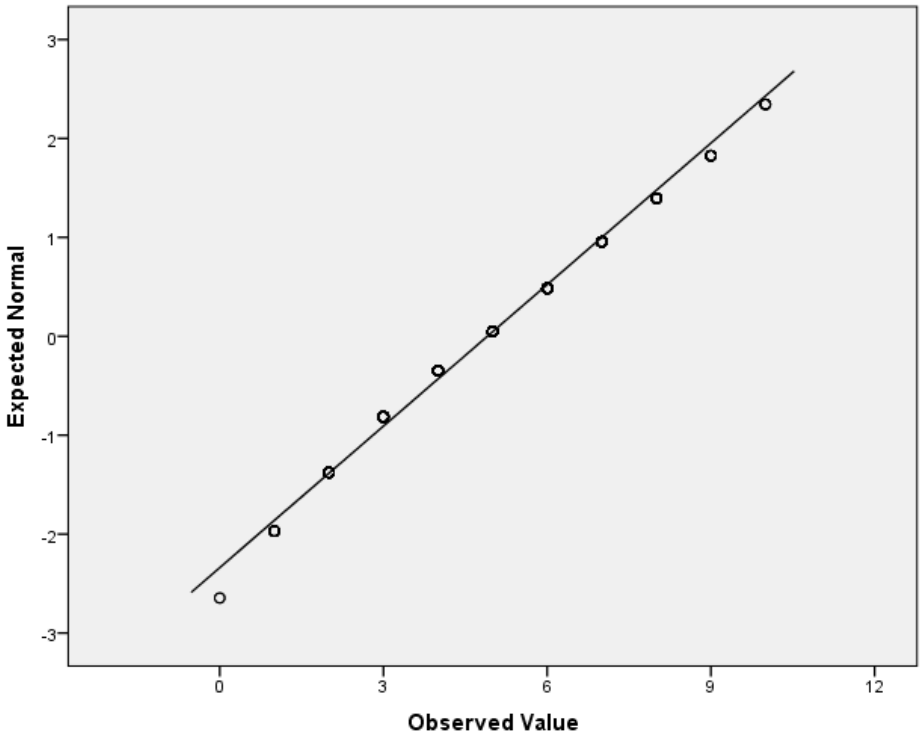

Figure 138: Q-Q Plot of Student Results in Foundations of Chemistry IA Exam 2015

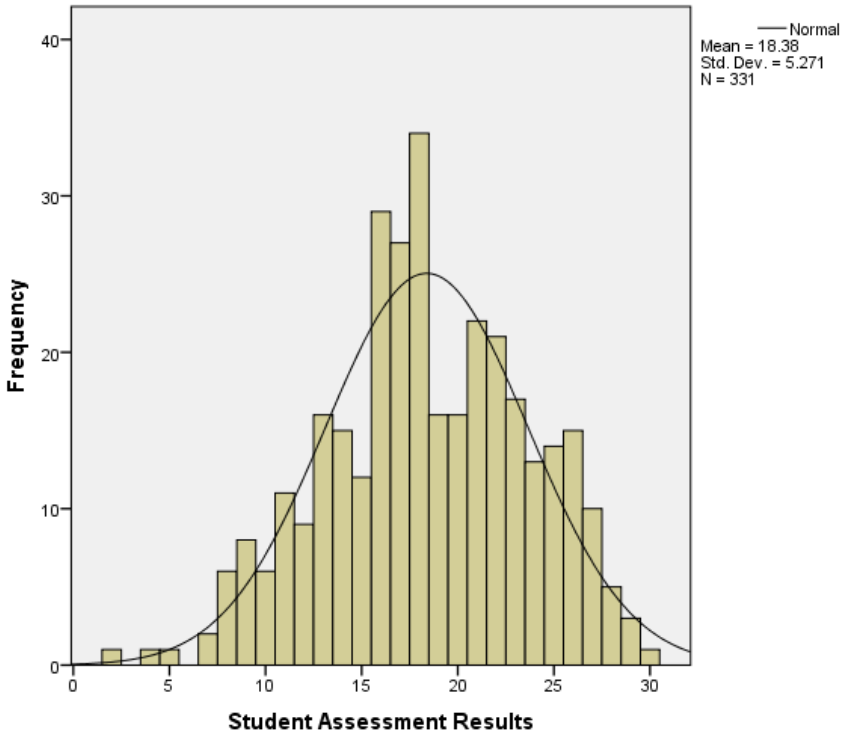

Figure 139: Student Scores Obtained in Foundations of Chemistry IA Redeemable Exam 2015

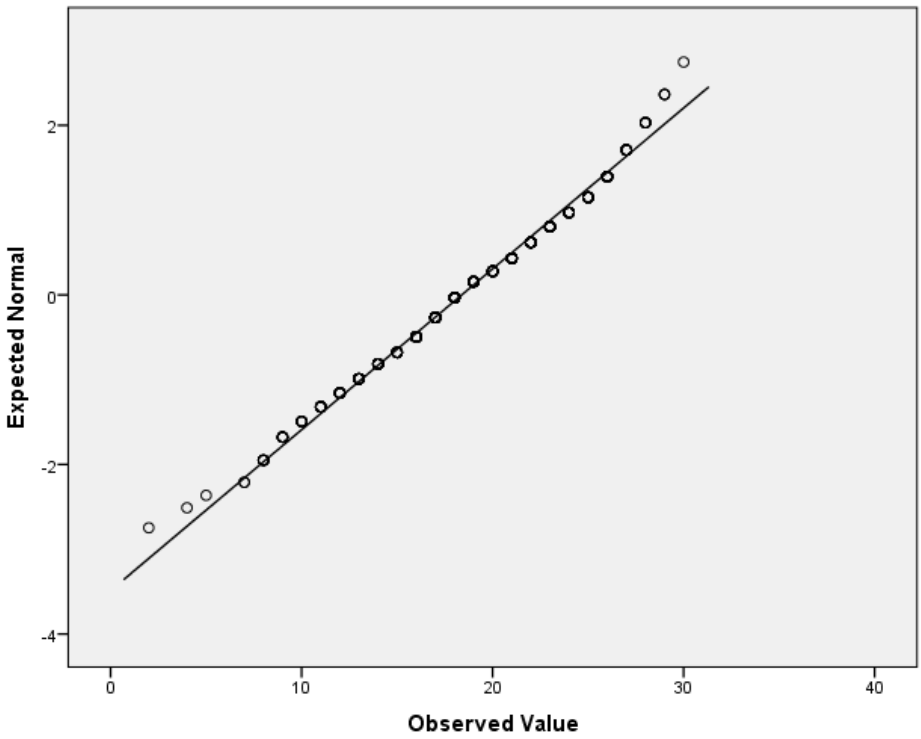

Figure 140: Q-Q Plot of Student Results in Foundations of Chemistry IA Redeemable Exam 2015

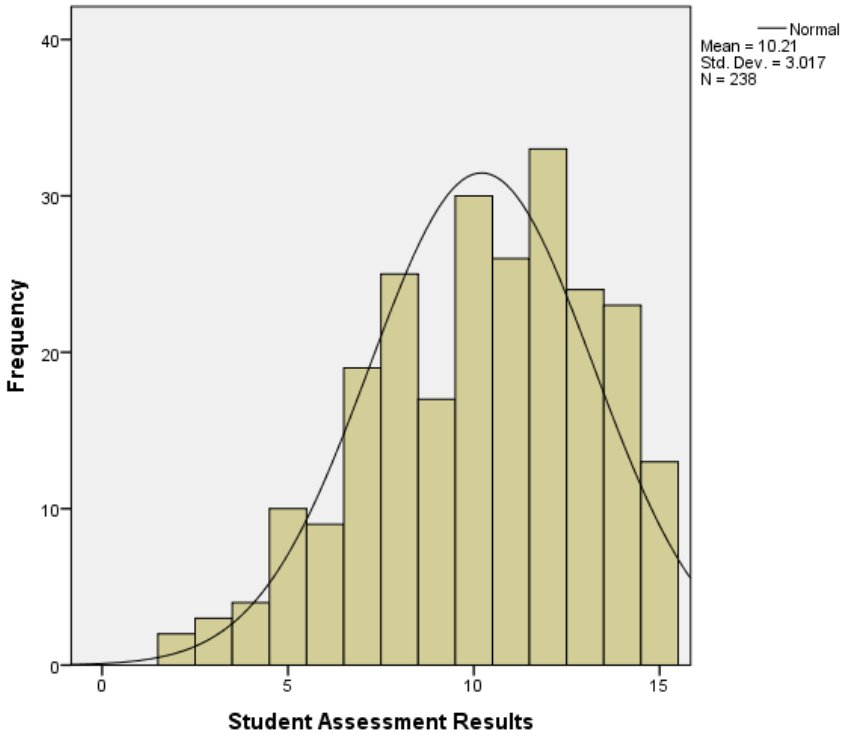

Figure 141: Student Scores Obtained in Foundations of Chemistry IB Lecture Test 1 2012

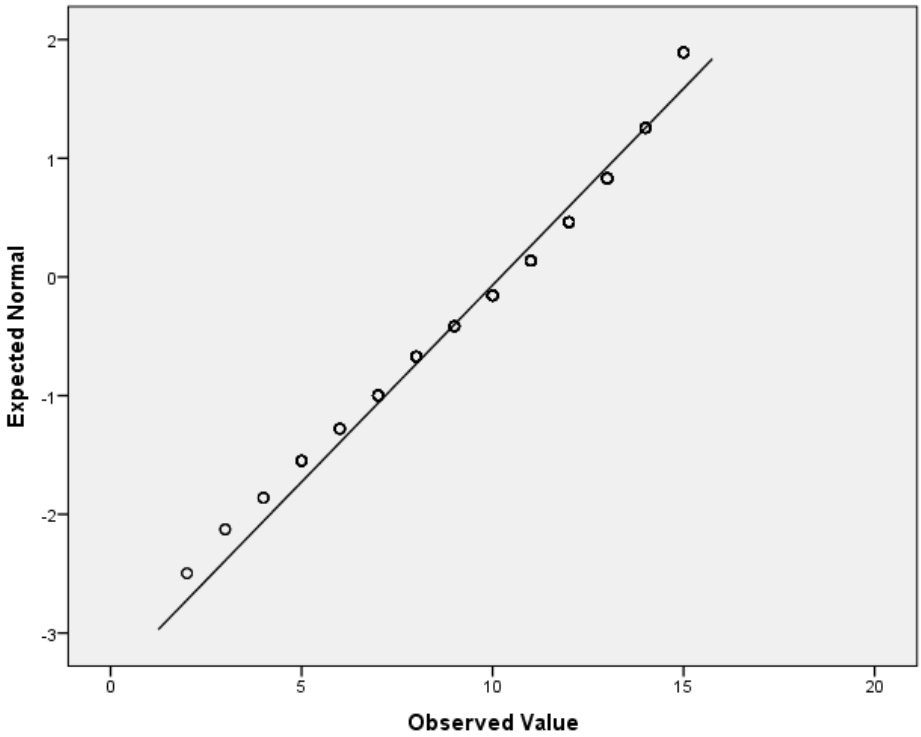

Figure 142: Q-Q Plot of Student Results in Foundations of Chemistry IB Lecture Test 1 2012

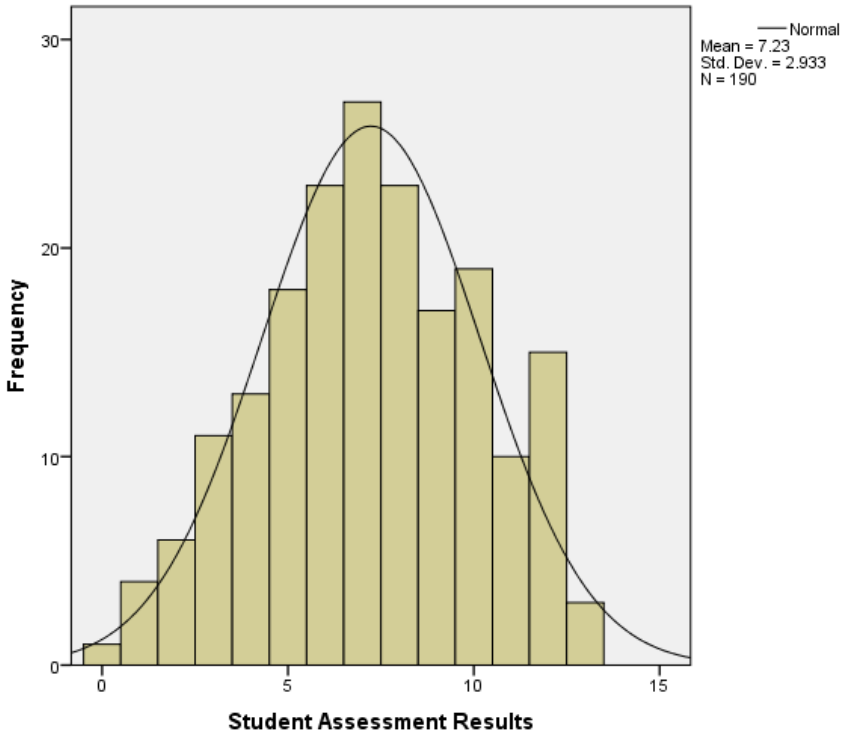

Figure 143: Student Scores Obtained in Foundations of Chemistry IB Lecture Test 2 2012

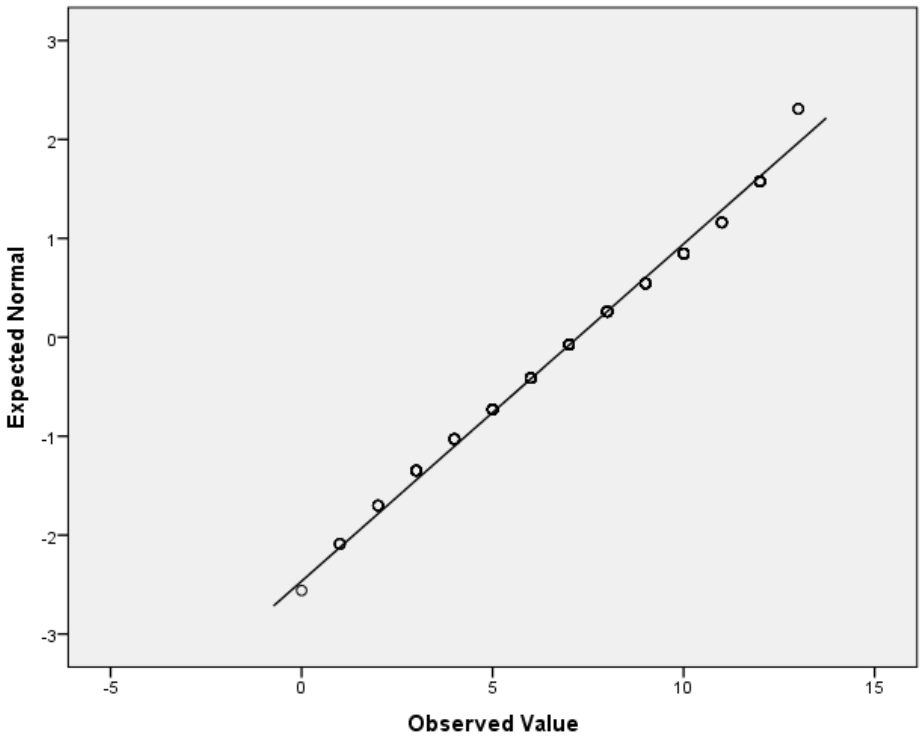

Figure 144: Q-Q Plot of Student Results in Foundations of Chemistry IB Lecture Test 2 2012

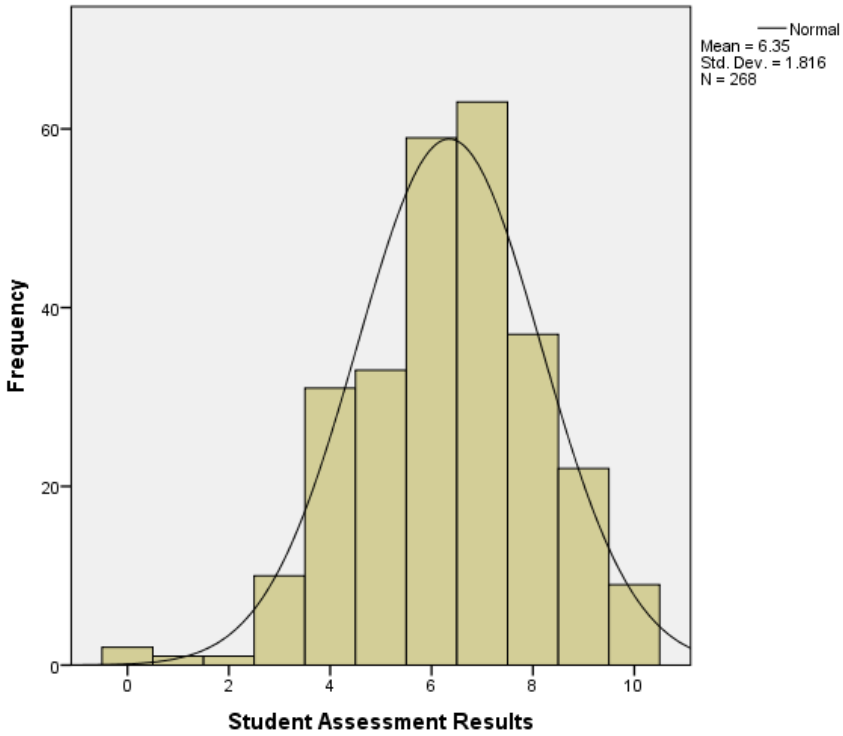

Figure 145: Student Scores Obtained in Foundations of Chemistry IB Exam 2012

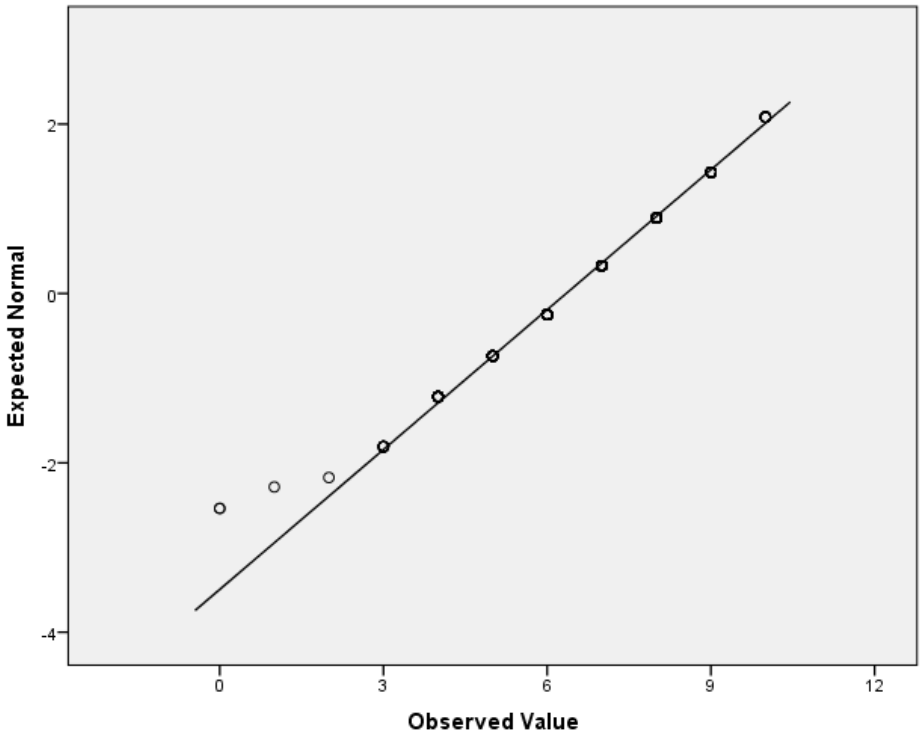

Figure 146: Q-Q Plot of Student Results in Foundations of Chemistry IB Exam 2012

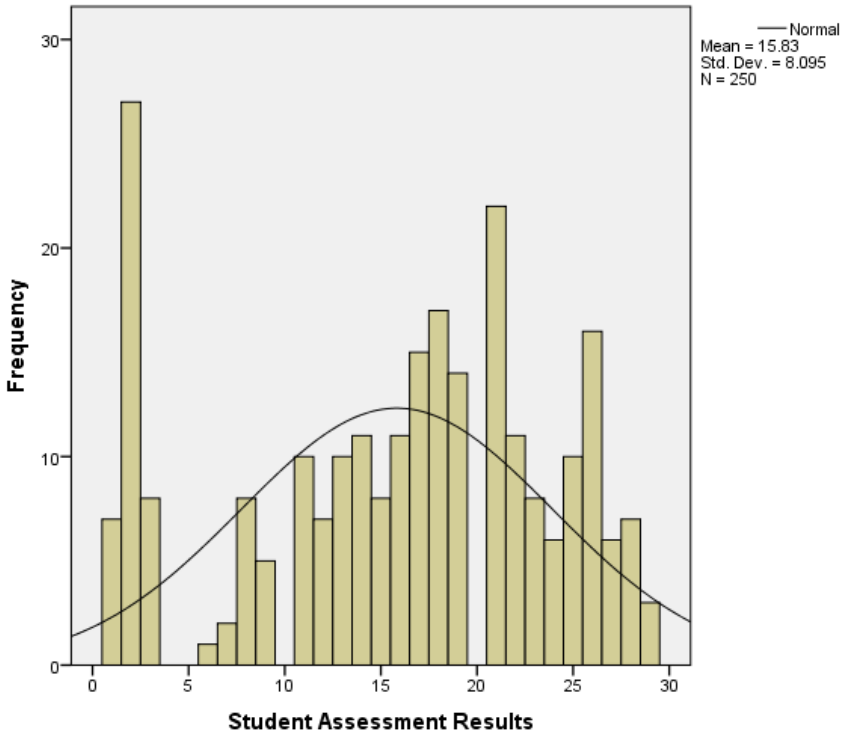

Figure 147: Student Scores Obtained in Foundations of Chemistry IB Redeemable Exam 2012

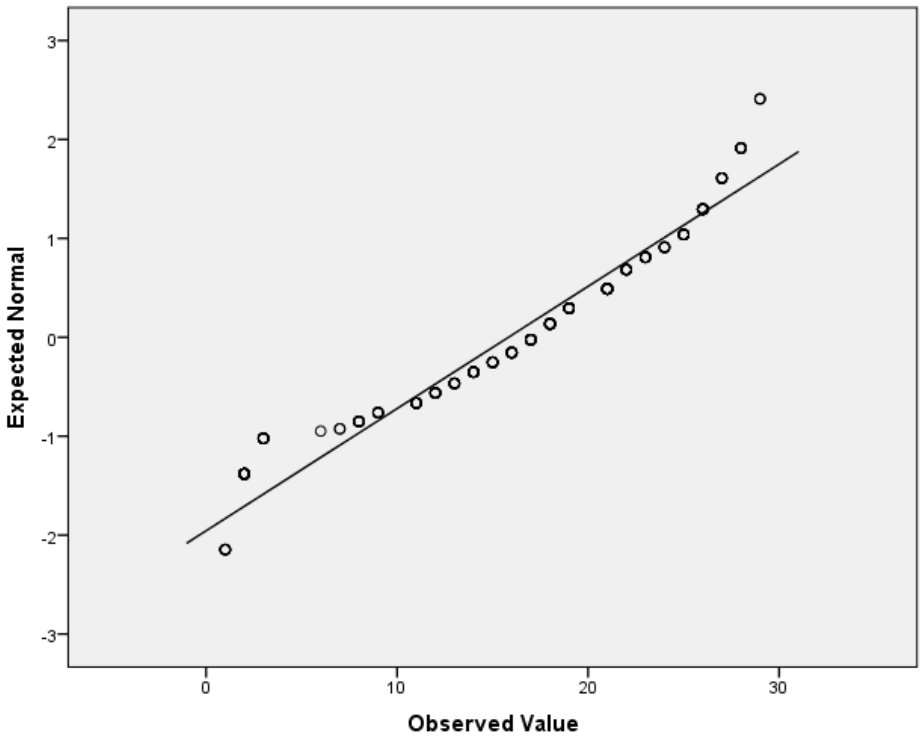

Figure 148: Q-Q Plot of Student Results in Foundations of Chemistry IB Redeemable Exam 2012

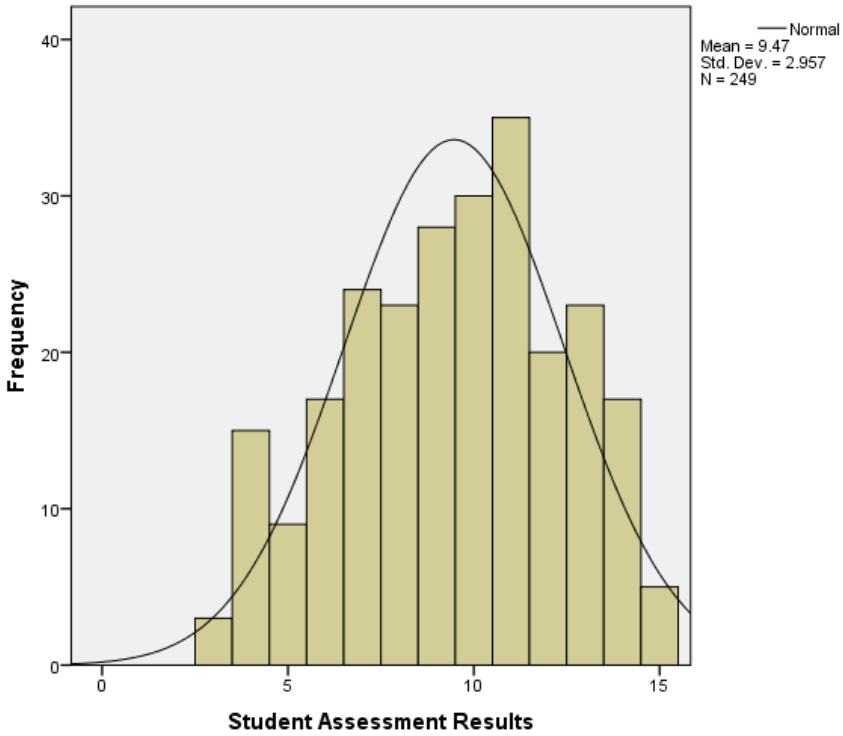

Figure 149: Student Scores Obtained in Foundations of Chemistry IB Lecture Test 1 2013

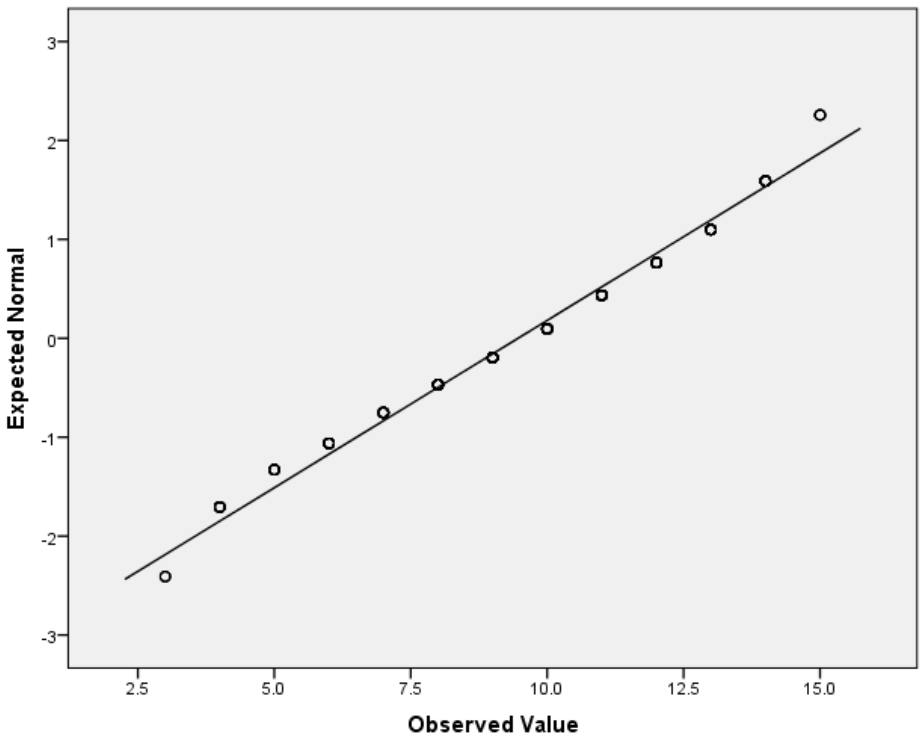

Figure 150: Q-Q Plot of Student Results in Foundations of Chemistry IB Lecture Test 1 2013

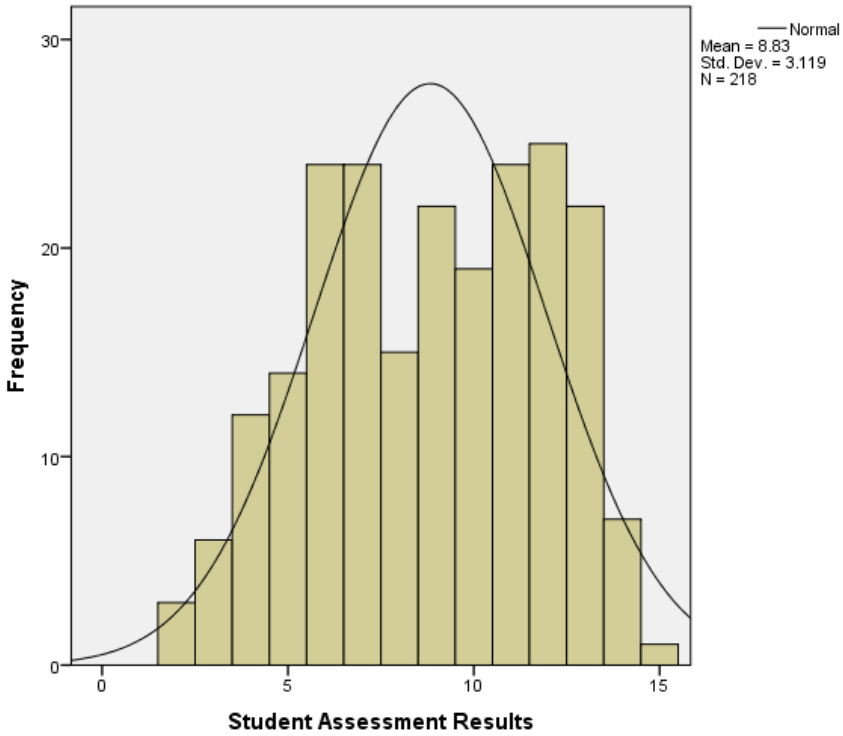

Figure 151: Student Scores Obtained in Foundations of Chemistry IB Lecture Test 2 2013

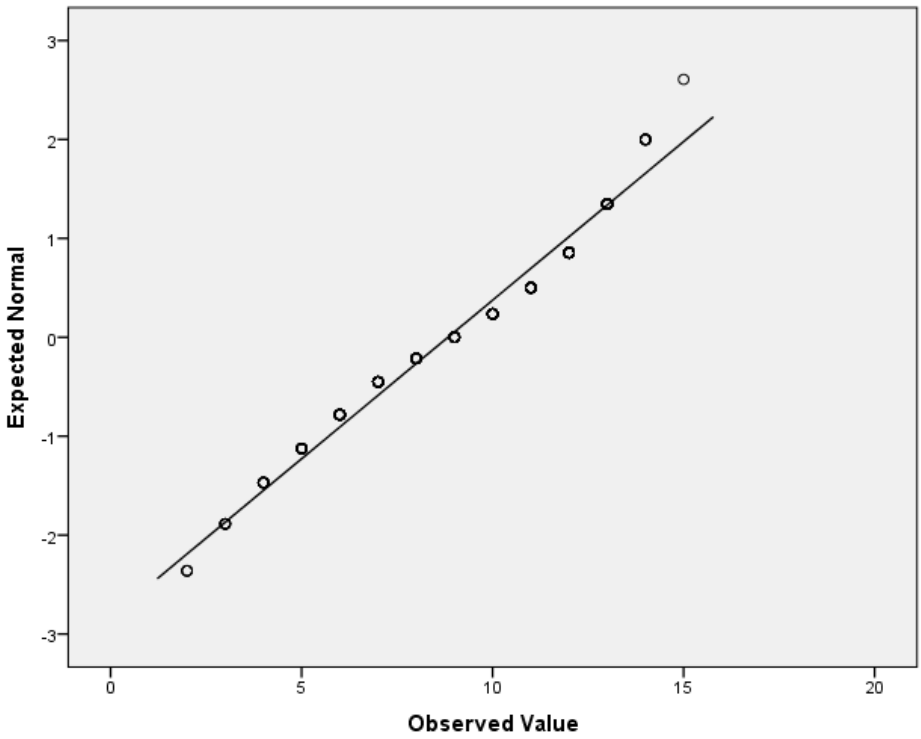

Figure 152: Q-Q Plot of Student Results in Foundations of Chemistry IB Lecture Test 2 2013

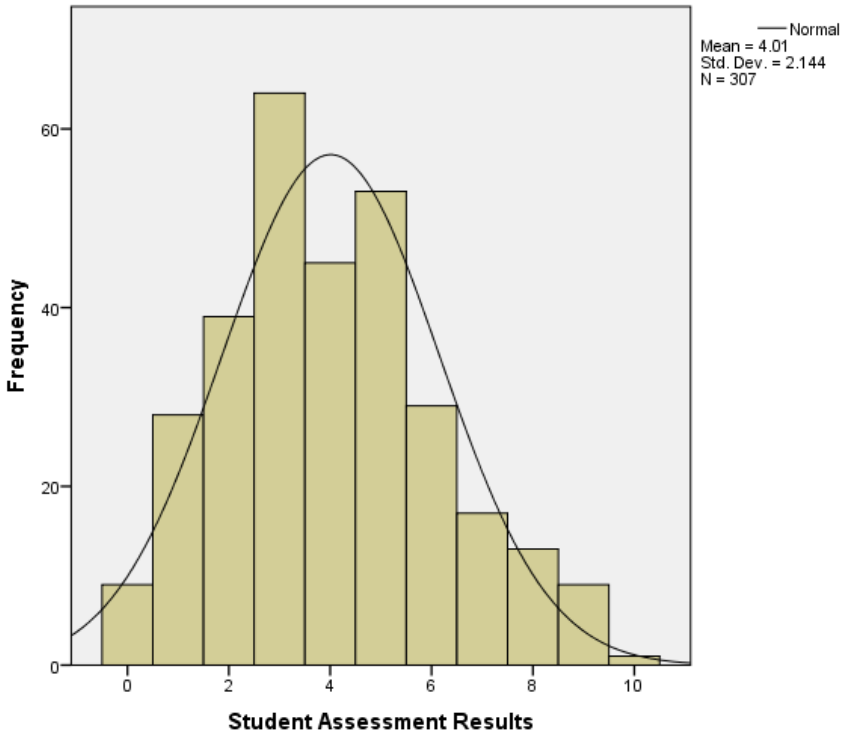

Figure 153: Student Scores Obtained in Foundations of Chemistry IB Exam 2013

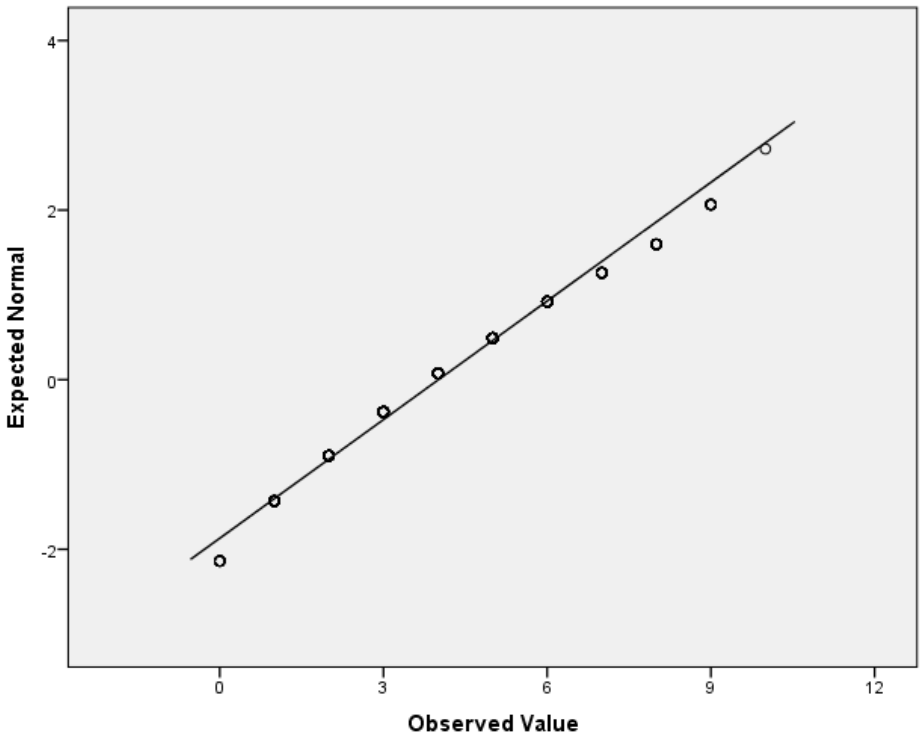

Figure 154: Q-Q Plot of Student Results in Foundations of Chemistry IB Exam 2013

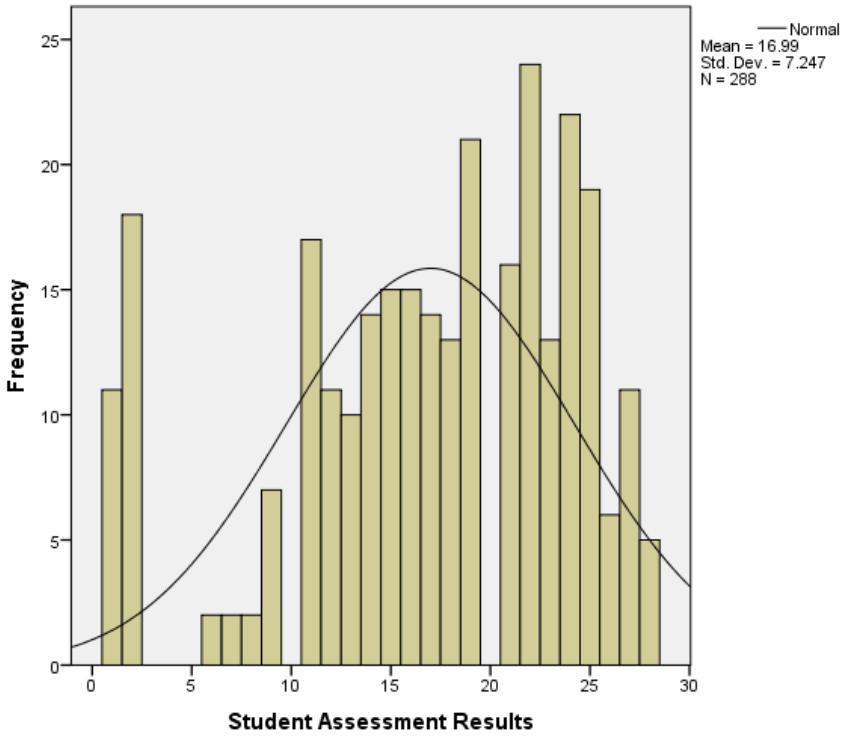

Figure 155: Student Scores Obtained in Foundations of Chemistry IB Redeemable Exam 2013

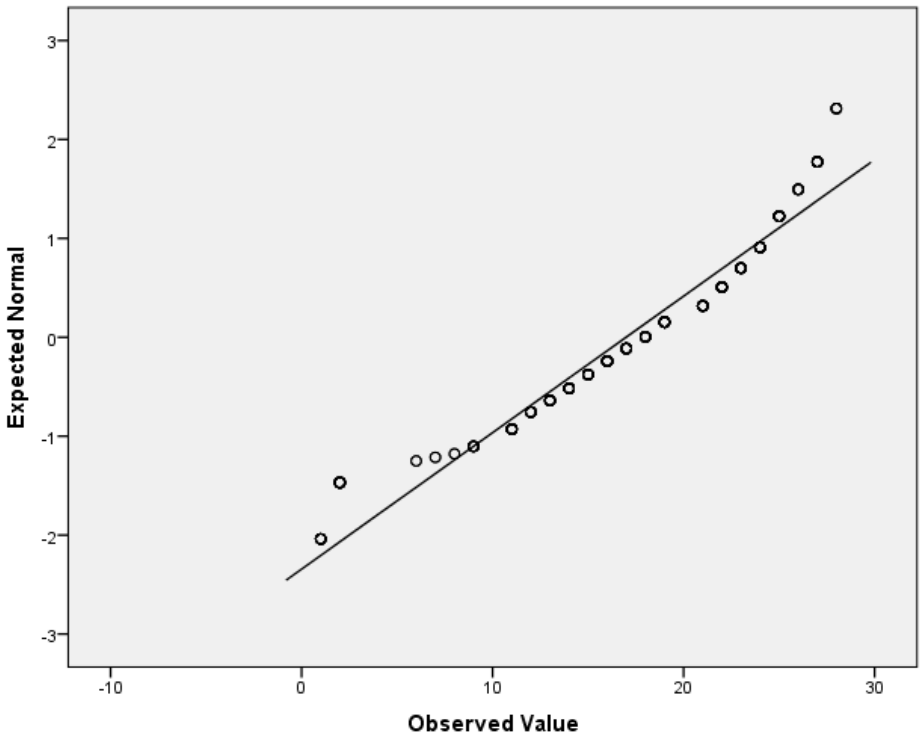

Figure 156: Q-Q Plot of Student Results in Foundations of Chemistry IB Redeemable Exam 2013

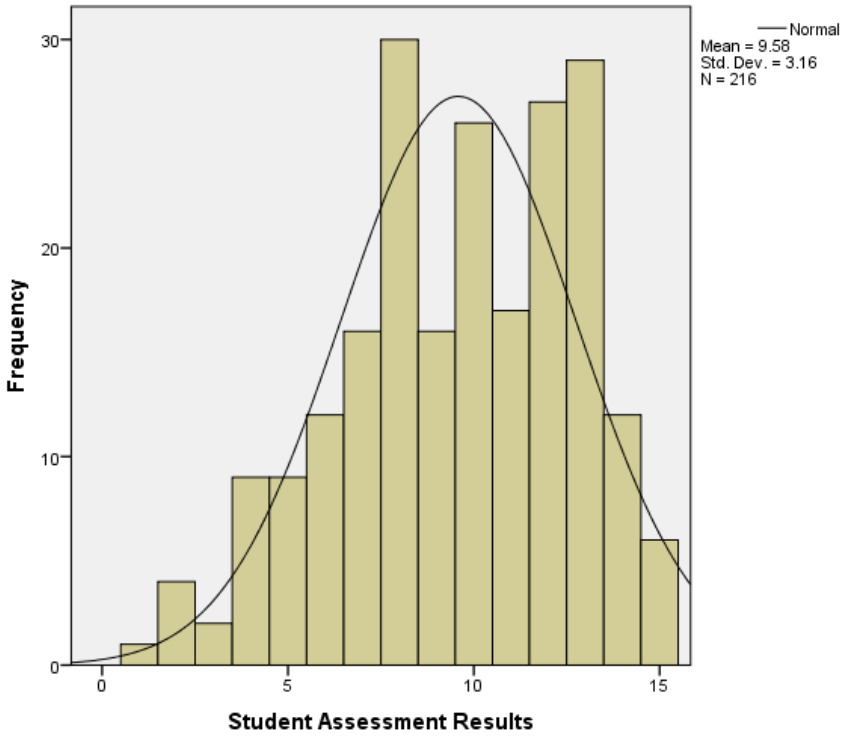

Figure 157: Student Scores Obtained in Foundations of Chemistry IB Lecture Test 1 2014

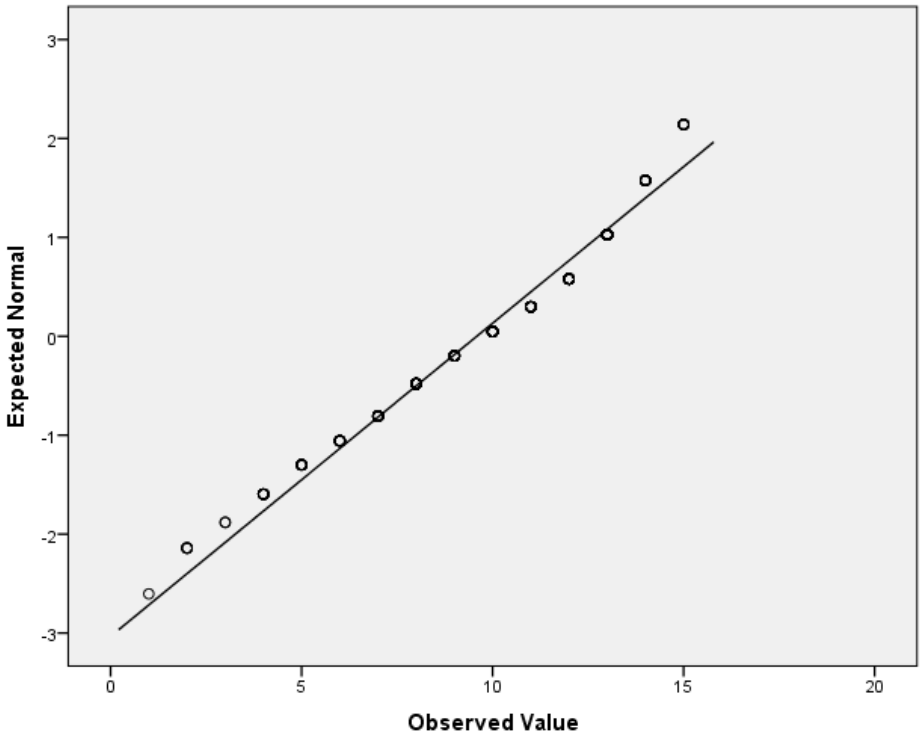

Figure 158: Q-Q Plot of Student Results in Foundations of Chemistry IB Lecture Test 1 2014

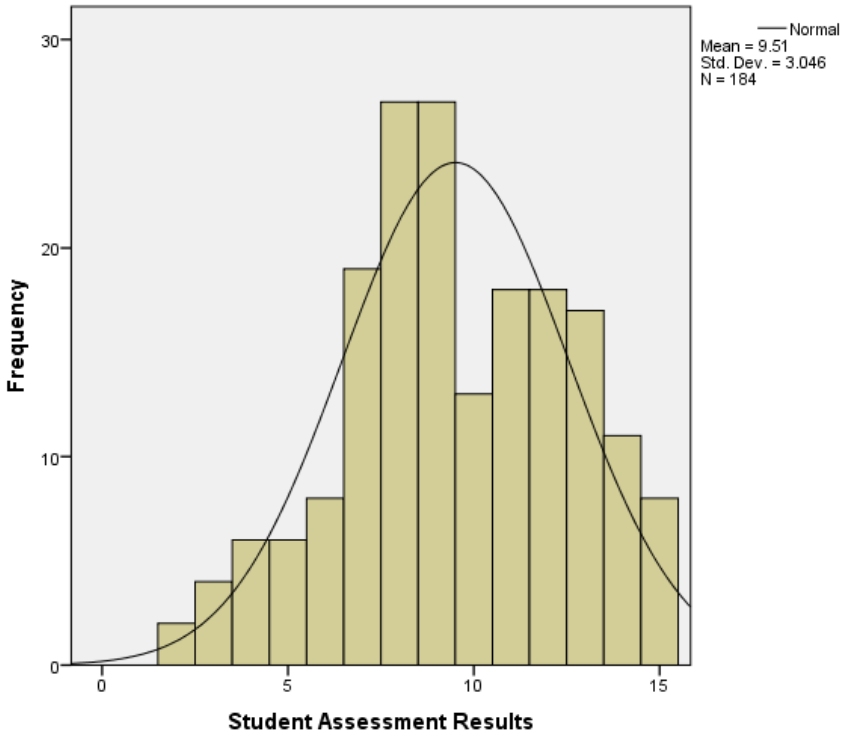

Figure 159: Student Scores Obtained in Foundations of Chemistry IB Lecture Test 2 2014

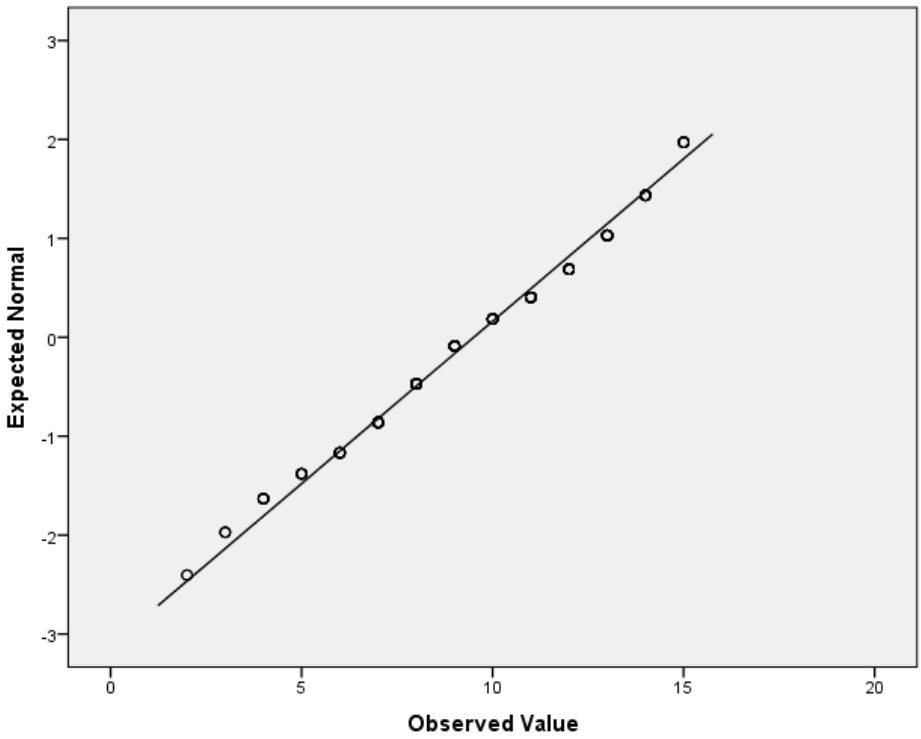

Figure 160: Q-Q Plot of Student Results in Foundations of Chemistry IB Lecture Test 2 2014

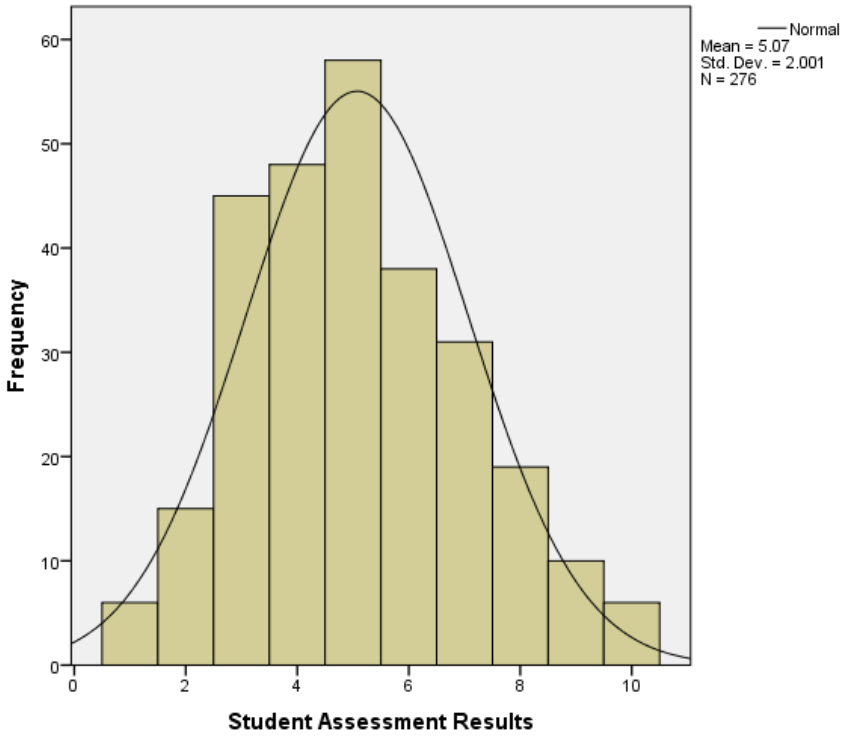

Figure 161: Student Scores Obtained in Foundations of Chemistry IB Exam 2014

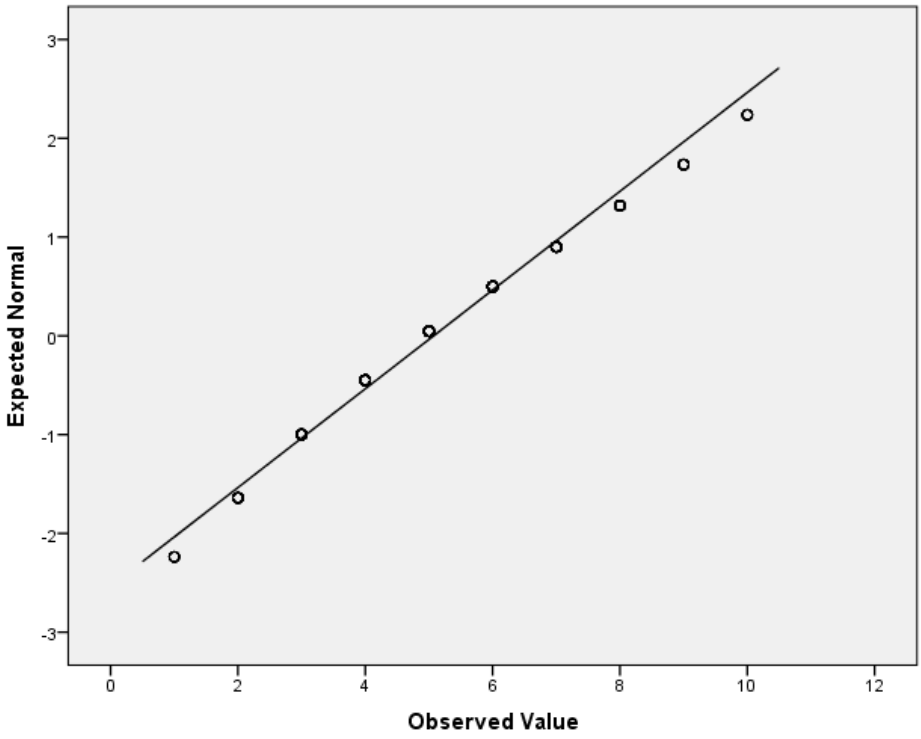

Figure 162: Q-Q Plot of Student Results in Foundations of Chemistry IB Exam 2014

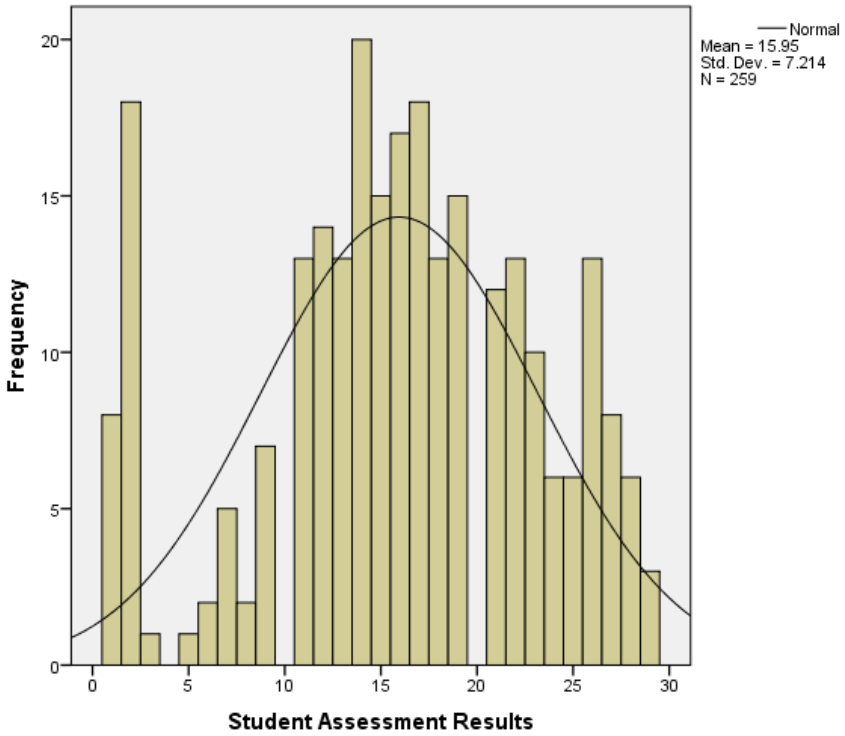

Figure 163: Student Scores Obtained in Foundations of Chemistry IB Redeemable Exam 2014

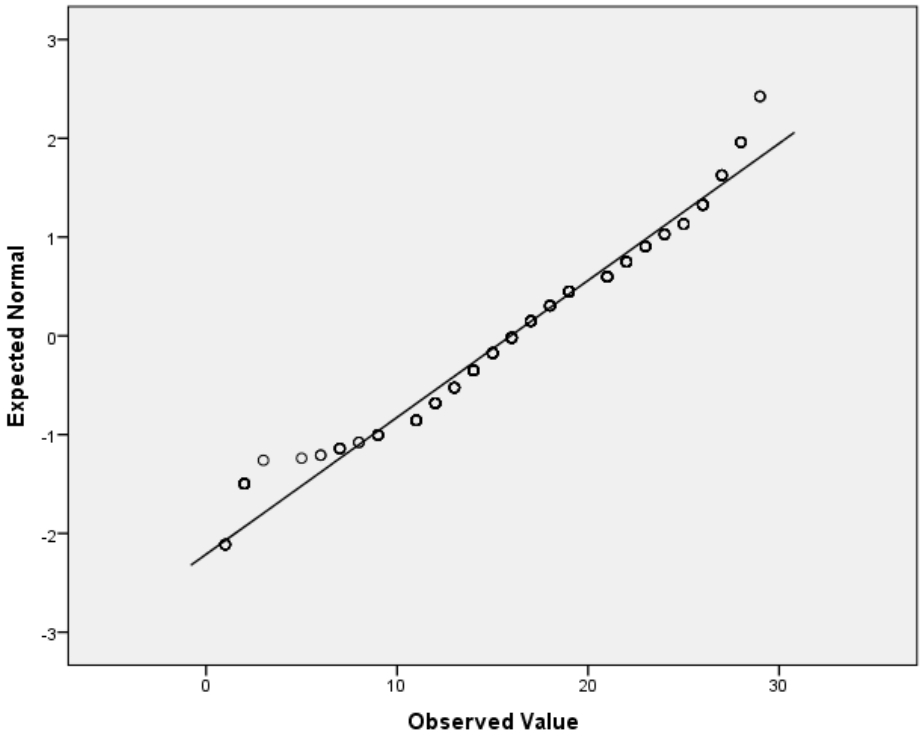

Figure 164: Q-Q Plot of Student Results in Foundations of Chemistry IB Redeemable Exam 2014

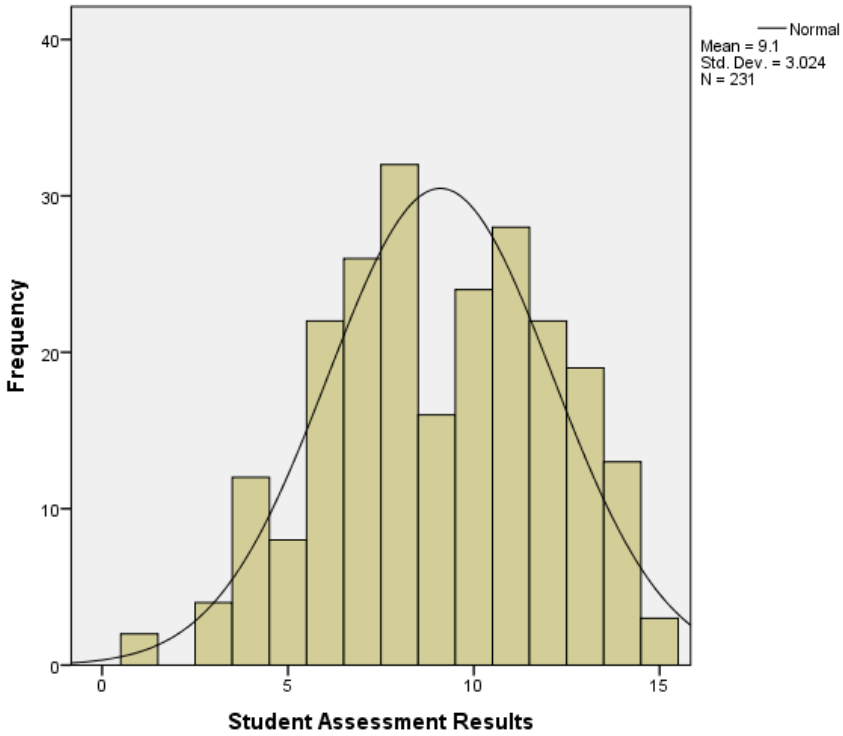

Figure 165: Student Scores Obtained in Foundations of Chemistry IB Lecture Test 1 2015

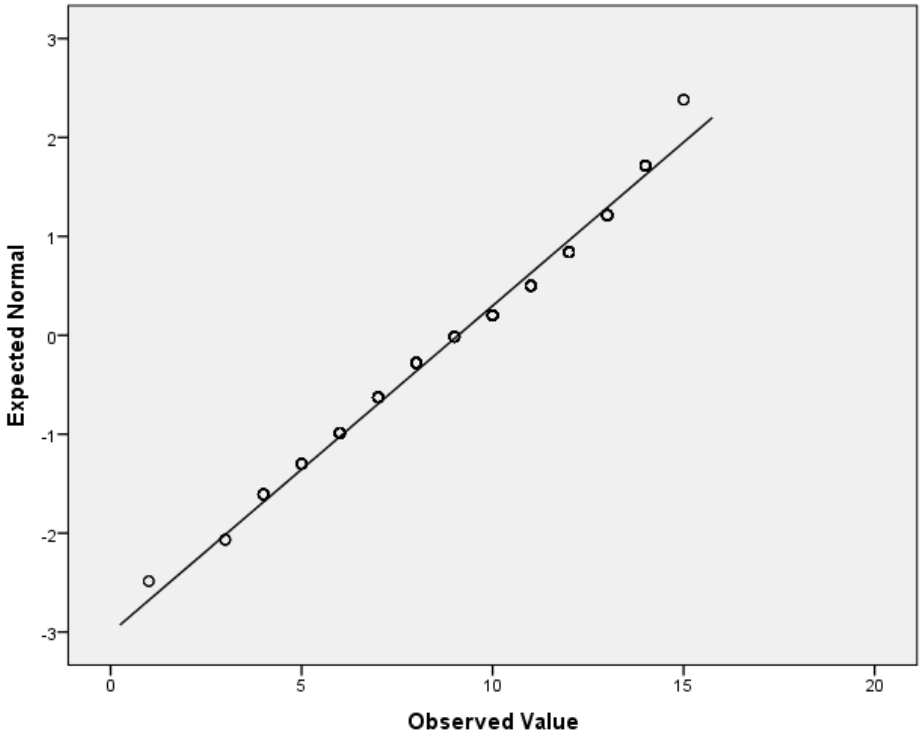

Figure 166: Q-Q Plot of Student Results in Foundations of Chemistry IB Lecture Test 1 2015

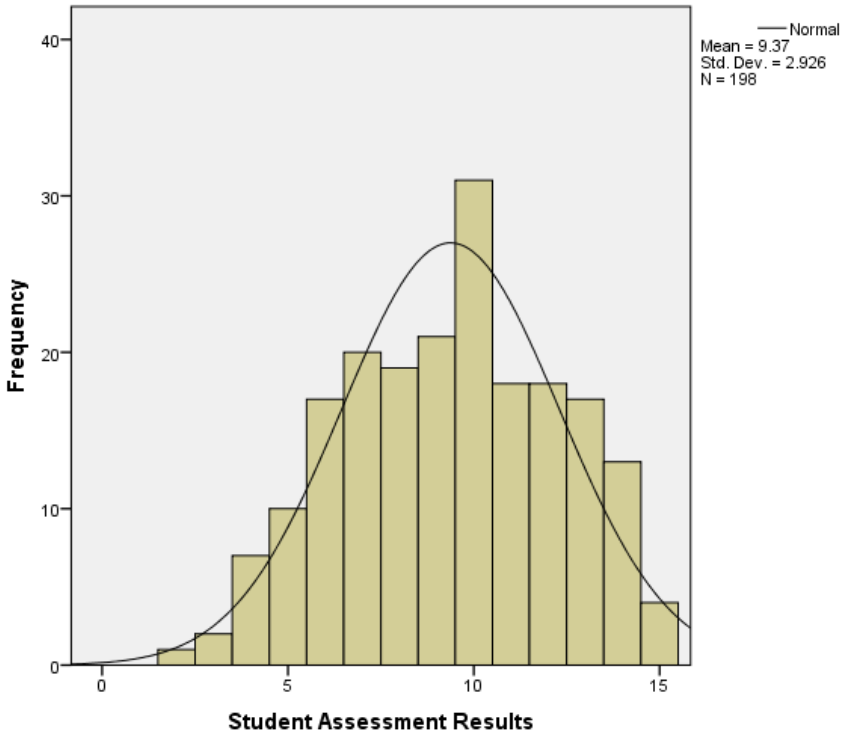

Figure 167: Student Scores Obtained in Foundations of Chemistry IB Lecture Test 2 2015

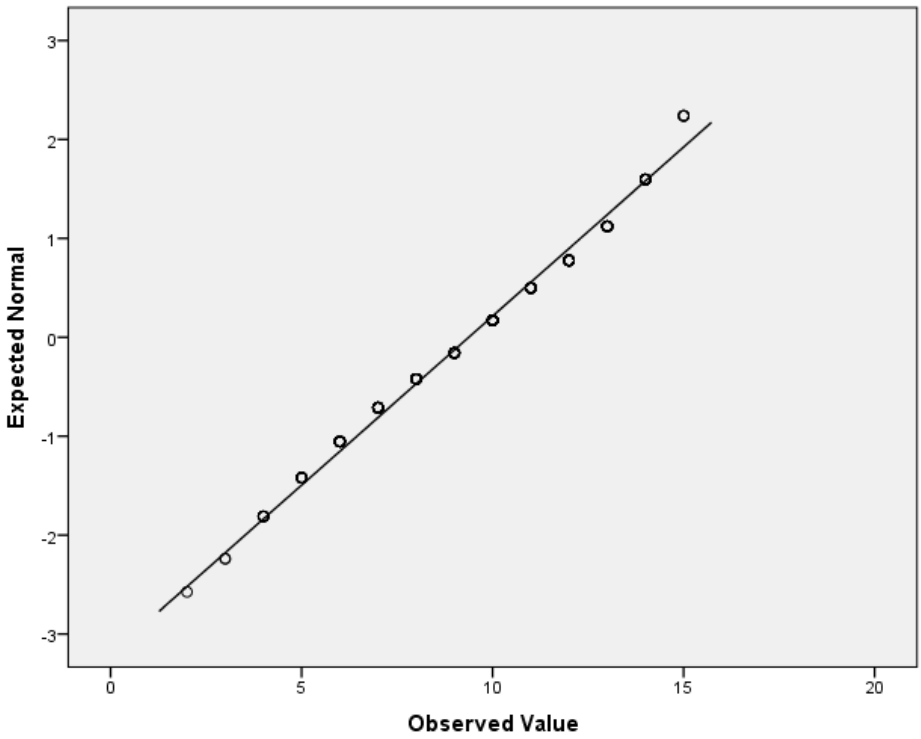

Figure 168: Q-Q Plot of Student Results in Foundations of Chemistry IB Lecture Test 2 2015

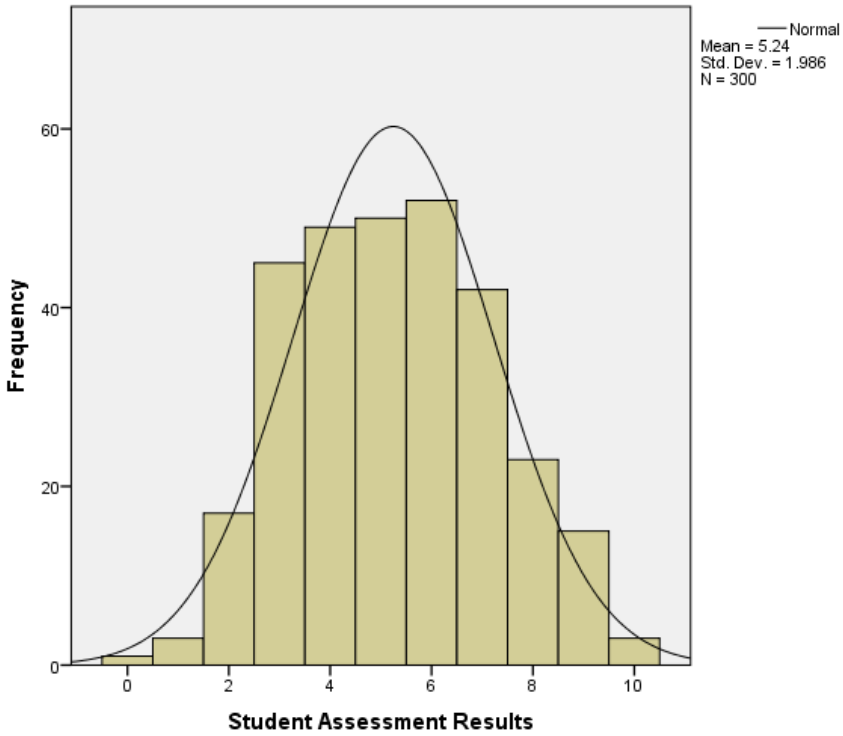

Figure 169: Student Scores Obtained in Foundations of Chemistry IB Exam 2015

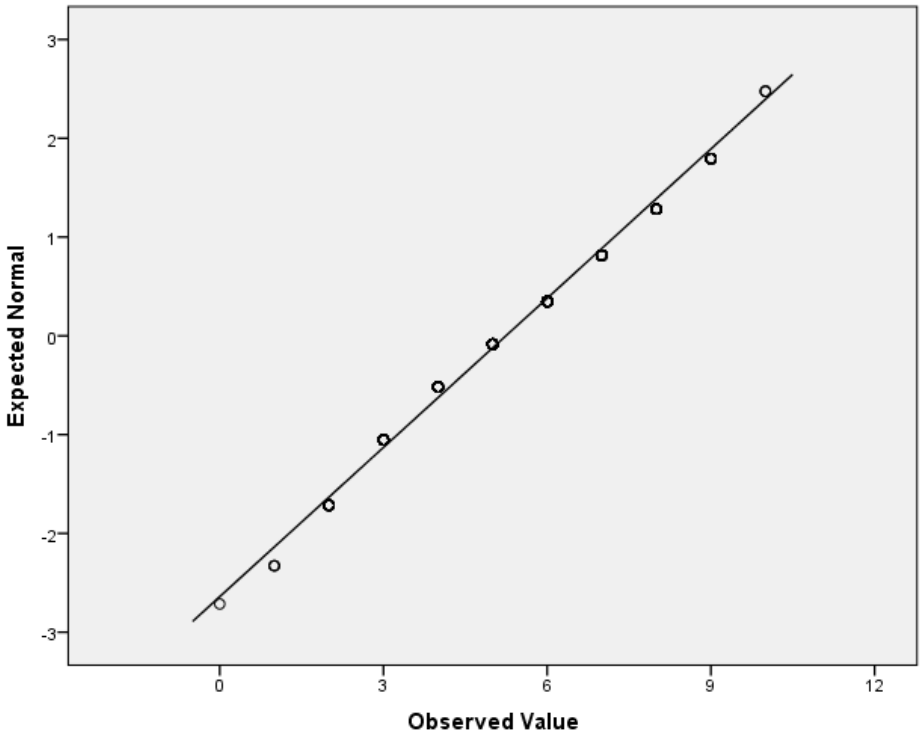

Figure 170: Q-Q Plot of Student Results in Foundations of Chemistry IB Exam 2015

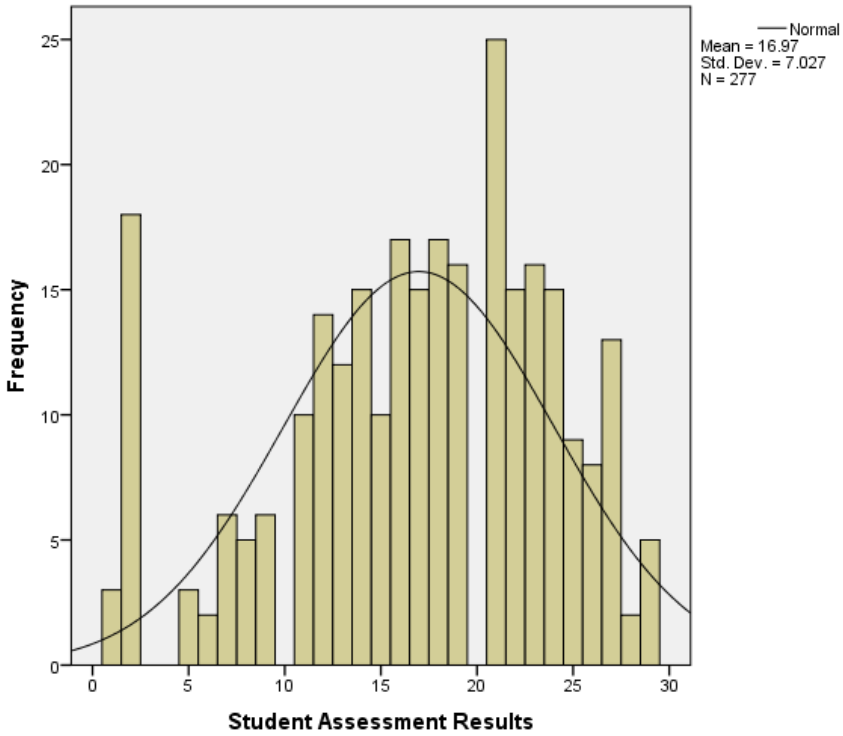

Figure 171: Student Scores Obtained in Foundations of Chemistry IB Redeemable Exam 2015

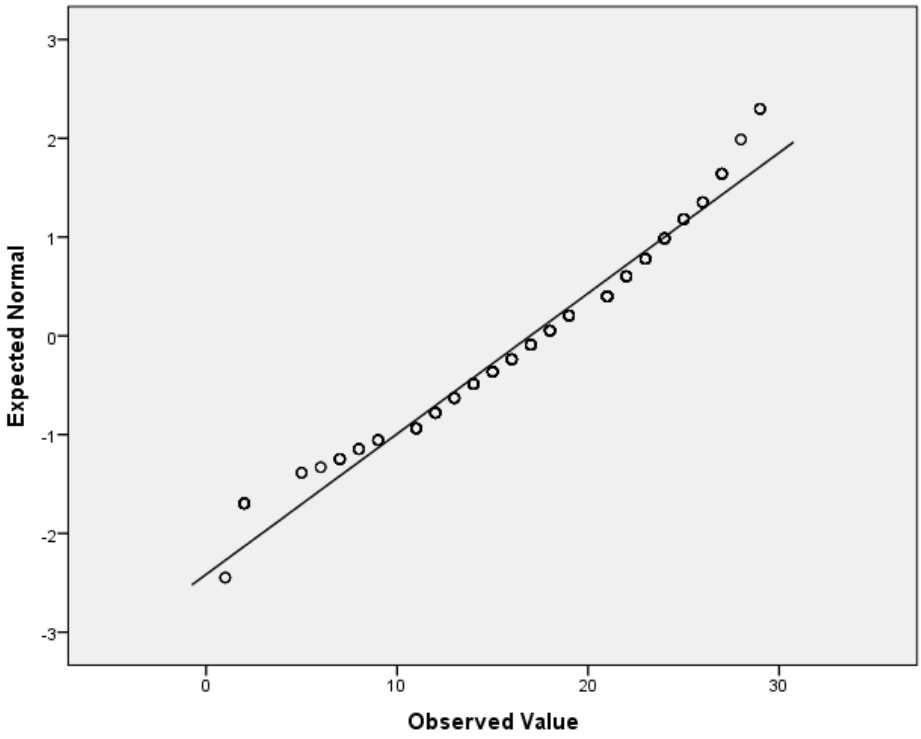

Figure 172: Q-Q Plot of Student Results in Foundations of Chemistry IB Redeemable Exam 2015

7.3 MCQ Assessment Task Evaluation and Analysis

Table 45: Evaluation of Each MCQ Assessment Task Undertaken within Chemistry IA Between 2012-2015 using CTT and Rasch Analysis

			CTT		Rasch Analysis				Rasch Dimensionality Residual Variance			
			KR-20	Ferguson's Delta	Student Reliability	Student Separation	Item Reliability	Item Separation	Measures Eigenvalue	Measure %	1st Contrast Eigenvalue	1st Contrast %
Chemistry IA	2012	Lecture Test 1	0.736	0.976	0.68	1.47	0.98	7.19	5.516	26.9	1.352	6.6
		Lecture Test 2	0.581	0.955	0.55	1.1	0.99	8.25	5.472	26.7	1.748	8.5
		Exam Test	0.538	0.954	0.5	1.01	0.98	7.46	3.432	25.5	1.524	15.2
		Redeemable Exam	0.828	0.982	0.82	2.12	0.99	8.43	11.516	27.7	1.790	4.3
	2013	Lecture Test 1	0.735	0.975	0.68	1.46	0.98	6.89	5.591	27.2	1.495	7.3
		Lecture Test 2	0.612	0.947	0.56	1.13	0.99	8.73	6.477	30.2	1.695	7.9
		Exam Test	0.511	0.952	0.48	0.96	0.98	7.09	3.202	24.3	1.632	12.4
		Redeemable Exam	0.882	0.986	0.85	2.41	0.99	8.3	11.701	27.5	1.814	4.3
	2014	Lecture Test 1	0.663	0.963	0.65	1.35	0.99	9.86	7.200	32.4	1.387	6.2
		Lecture Test 2	0.638	0.956	0.6	1.22	0.98	6.64	4.910	24.7	1.617	8.1
		Exam Test	0.609	0.965	0.53	1.07	0.98	6.95	3.726	27.1	1.371	9.5
		Redeemable Exam	0.847	0.984	0.84	2.26	0.98	7.83	10.214	25.4	1.591	4
	2015	Lecture Test 1	0.662	0.960	0.64	1.34	0.99	9.95	6.934	31.6	1.390	6.3
		Lecture Test 2	0.646	0.962	0.6	1.23	0.98	6.42	4.533	23.2	1.366	7
		Exam Test	0.604	0.963	0.53	1.07	0.98	7.06	3.435	25.6	1.329	9.9
		Redeemable Exam	0.851	0.978	0.84	2.29	0.99	8.88	10.413	25.8	1.753	4.3

Table 46: Evaluation of Each MCQ Assessment Task Undertaken within Chemistry IB Between 2012-2015 using CTT and Rasch Analysis

			CTT		Rasch Analysis				Rasch Dimensionality Residual Variance			
			KR-20	Ferguson's Delta	Student Reliability	Student Separation	Item Reliability	Item Separation	Measures Eigenvalue	Measure %	1st Contrast Eigenvalue	1st Contrast %
Chemistry IB	2012	Lecture Test 1	0.643	0.966	0.62	1.29	0.96	4.87	3.793	20.2	1.511	8.0
		Lecture Test 2	0.702	0.969	0.69	1.48	0.97	5.41	4.645	23.6	1.537	7.8
		Exam Test	0.456	0.933	0.42	0.86	0.99	8.11	3.585	28.5	1.318	10.5
		Redeemable Exam	0.840	0.981	0.79	1.92	0.98	6.36	8.437	21.9	1.651	4.3
	2013	Lecture Test 1	0.656	0.966	0.62	1.27	0.96	4.71	3.761	20	1.621	8.6
		Lecture Test 2	0.674	0.967	0.65	1.36	0.97	5.84	4.971	24.9	1.497	7.5
		Exam Test	0.544	0.952	0.51	1.02	0.99	9.02	4.829	32.6	1.279	8.6
		Redeemable Exam	0.875	0.985	0.81	2.05	0.97	5.93	8.843	22.8	1.831	4.7
	2014	Lecture Test 1	0.725	0.976	0.67	1.41	0.96	5.08	4.377	22.6	1.574	8.1
		Lecture Test 2	0.696	0.971	0.67	1.43	0.97	6.16	4.994	25	1.515	7.6
		Exam Test	0.567	0.956	0.54	1.08	0.99	8.69	4.254	29.8	1.376	9.6
		Redeemable Exam	0.884	0.983	0.8	1.98	0.97	5.88	7.866	20.8	1.615	4.3
	2015	Lecture Test 1	0.677	0.971	0.65	1.35	0.97	5.53	4.153	21.7	1.440	7.5
		Lecture Test 2	0.647	0.962	0.63	1.32	0.98	7.6	5.321	26.2	1.383	6.8
		Exam Test	0.553	0.959	0.52	1.03	0.98	7.68	3.802	27.5	1.361	9.9
		Redeemable Exam	0.870	0.985	0.81	2.07	0.98	6.66	9.011	23.1	1.646	4.2

Table 47: Evaluation of Each MCQ Assessment Task Undertaken within Foundations of Chemistry IA Between 2012-2015 using CTT and Rasch Analysis

			CTT		Rasch Analysis				Rasch Dimensionality Residual Variance			
			KR-20	Ferguson's Delta	Student Reliability	Student Separation	Item Reliability	Item Separation	Measures Eigenvalue	Measure %	1st Contrast Eigenvalue	1st Contrast %
Foundations of Chemistry IA	2012	Lecture Test 1	0.623	0.934	0.49	0.98	0.97	5.67	5.519	26.9	1.514	7.4
		Lecture Test 2	0.689	0.959	0.65	1.36	0.99	8.39	9.520	38.8	1.532	6.2
		Exam Test	0.549	0.950	0.52	1.04	0.98	7.05	4.278	30.0	1.450	10.2
		Redeemable Exam	0.803	0.970	0.78	1.86	0.98	7.64	14.375	32.4	1.825	4.1
	2013	Lecture Test 1	0.632	0.937	0.52	1.05	0.98	7.00	7.375	33.0	1.552	6.9
		Lecture Test 2	0.707	0.966	0.68	1.46	0.99	8.25	9.716	39.3	1.552	6.3
		Exam Test	0.574	0.955	0.54	1.08	0.98	8.07	4.634	31.7	1.440	9.8
		Redeemable Exam	0.899	0.975	0.85	2.42	0.99	8.77	14.536	32.6	1.713	3.8
	2014	Lecture Test 1	0.712	0.941	0.56	1.13	0.97	6.00	7.516	33.4	1.506	6.7
		Lecture Test 2	0.649	0.952	0.63	1.29	0.98	7.75	9.366	38.4	1.528	6.3
		Exam Test	0.583	0.956	0.56	1.12	0.98	8.05	5.010	33.4	1.517	10.1
		Redeemable Exam	0.906	0.979	0.81	2.04	0.99	8.33	15.684	34.3	1.822	4.0
	2015	Lecture Test 1	0.712	0.949	0.59	1.19	0.98	6.77	7.628	33.7	1.572	6.9
		Lecture Test 2	0.717	0.960	0.68	1.46	0.98	7.75	9.891	39.7	1.634	6.6
		Exam Test	0.572	0.955	0.54	1.08	0.99	8.44	4.890	32.8	1.426	9.6
		Redeemable Exam	0.914	0.976	0.80	2.03	0.99	8.66	15.115	33.5	1.629	3.6

Table 48: Evaluation of Each MCQ Assessment Task Undertaken within Foundations of Chemistry IB Between 2012-2015 using CTT and Rasch Analysis

			CTT		Rasch Analysis				Rasch Dimensionality Residual Variance			
			KR-20	Ferguson's Delta	Student Reliability	Student Separation	Item Reliability	Item Separation	Measures Eigenvalue	Measure %	1st Contrast Eigenvalue	1st Contrast %
Foundations of Chemistry IB	2012	Lecture Test 1	0.751	0.965	0.68	1.46	0.98	6.66	8.973	36.8	1.545	6.5
		Lecture Test 2	0.709	0.965	0.68	1.47	0.98	6.42	8.136	36.8	1.505	6.8
		Exam Test	0.513	0.923	0.42	0.84	0.99	8.38	6.314	38.7	1.447	8.9
		Redeemable Exam	0.913	0.981	0.83	2.23	0.98	7.13	15.074	33.4	1.769	3.9
	2013	Lecture Test 1	0.717	0.966	0.67	1.42	0.98	7.33	8.409	35.9	1.474	6.3
		Lecture Test 2	0.726	0.969	0.68	1.45	0.97	6.11	7.400	33.0	1.525	6.8
		Exam Test	0.595	0.952	0.51	1.03	0.98	6.63	4.332	30.2	1.367	9.5
		Redeemable Exam	0.843	0.977	0.80	2.02	0.98	7.19	13.827	33.1	1.601	3.8
	2014	Lecture Test 1	0.753	0.965	0.69	1.50	0.97	6.24	8.118	35.1	1.559	6.7
		Lecture Test 2	0.742	0.964	0.70	1.54	0.98	6.40	9.092	37.7	1.691	7.0
		Exam Test	0.533	0.943	0.50	1.00	0.98	7.57	5.115	33.8	1.561	10.3
		Redeemable Exam	0.896	0.983	0.83	2.18	0.98	7.38	14.879	33.2	1.879	4.2
	2015	Lecture Test 1	0.716	0.966	0.68	1.44	0.98	6.71	7.642	33.8	1.654	7.3
		Lecture Test 2	0.712	0.965	0.68	1.44	0.98	6.73	8.651	36.6	1.674	7.1
		Exam Test	0.521	0.948	0.49	0.97	0.98	7.75	4.954	33.1	1.388	9.3
		Redeemable Exam	0.907	0.982	0.81	2.10	0.98	7.09	13.368	30.8	1.845	4.3

7.4 MCQ Assessment Wright Maps

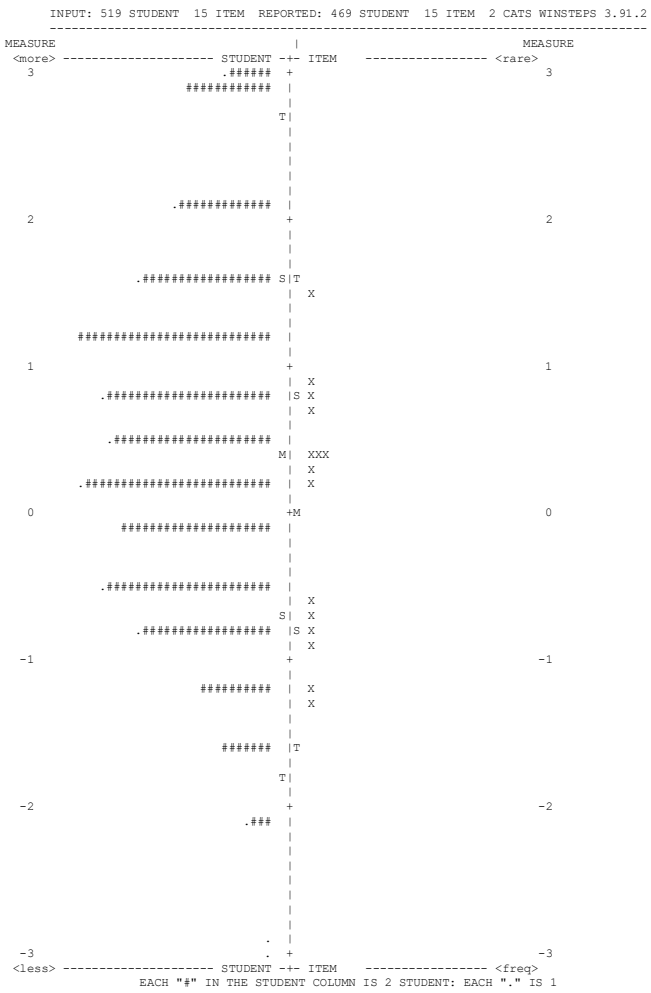

Figure 173: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2012 Lecture Test 1

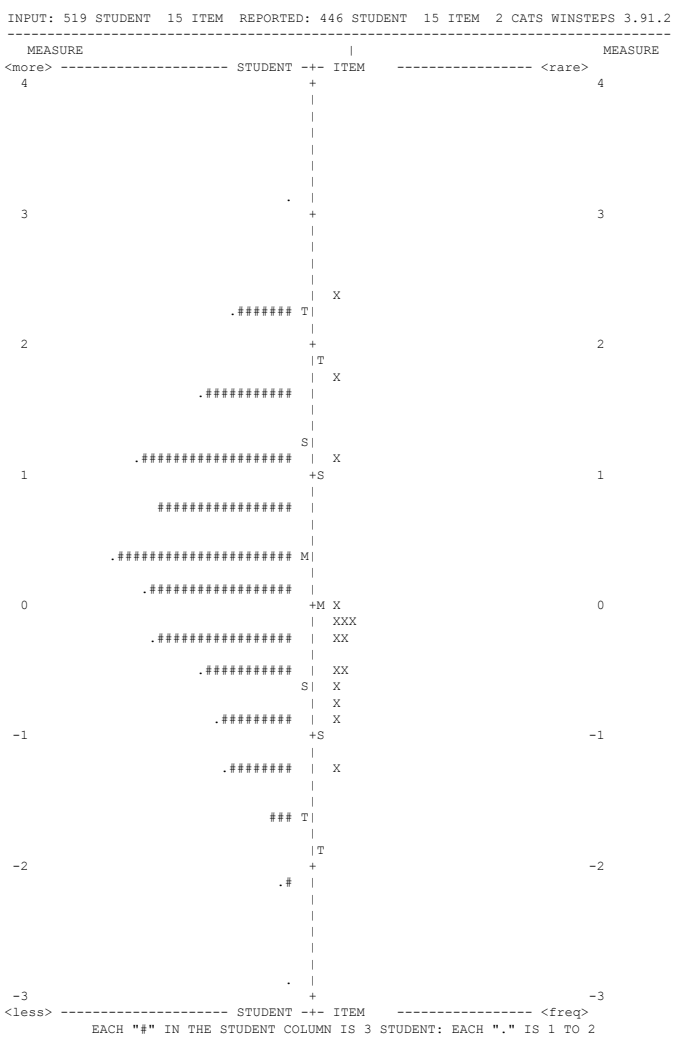

Figure 174: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2012 Lecture Test 2

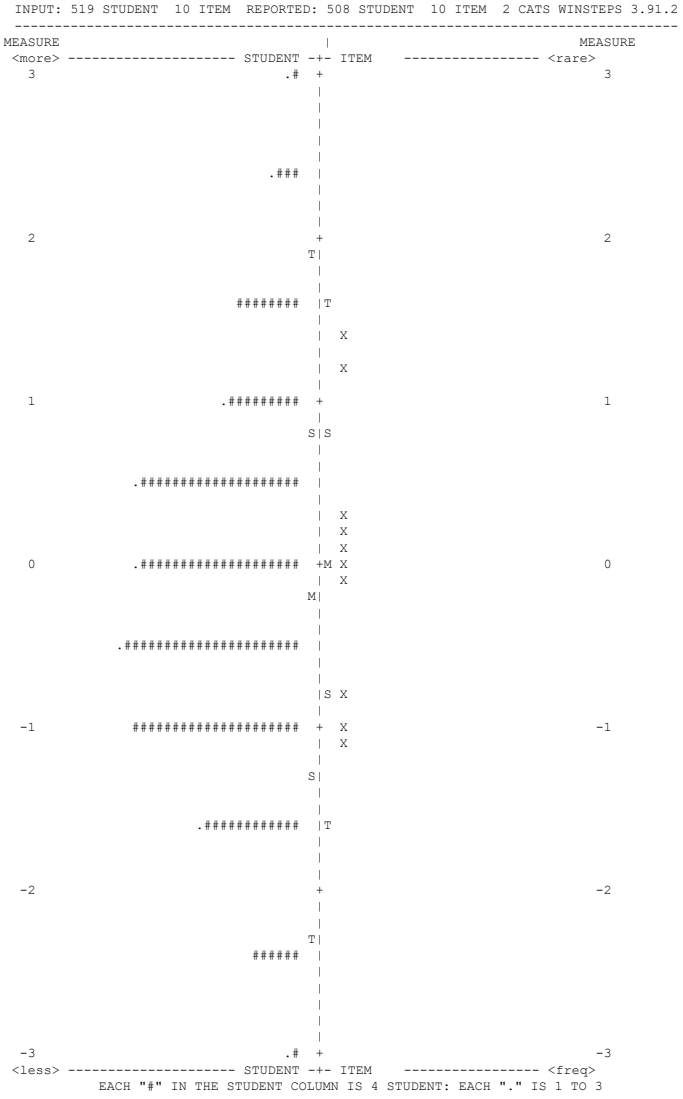

Figure 175: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2012 Exam

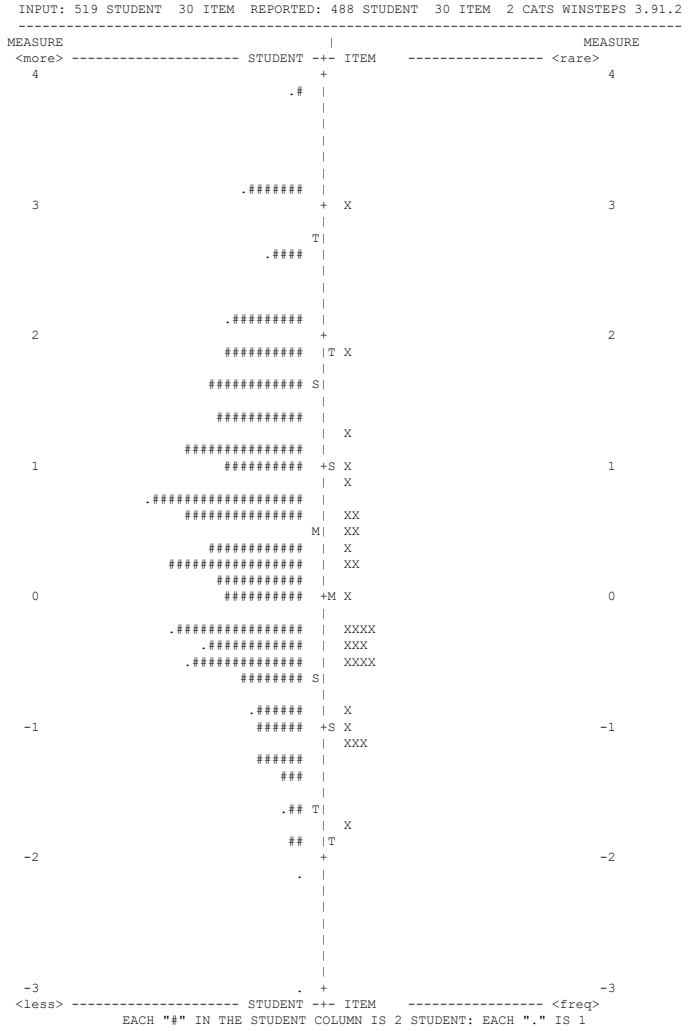

Figure 176: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2012 Redeemable Exam

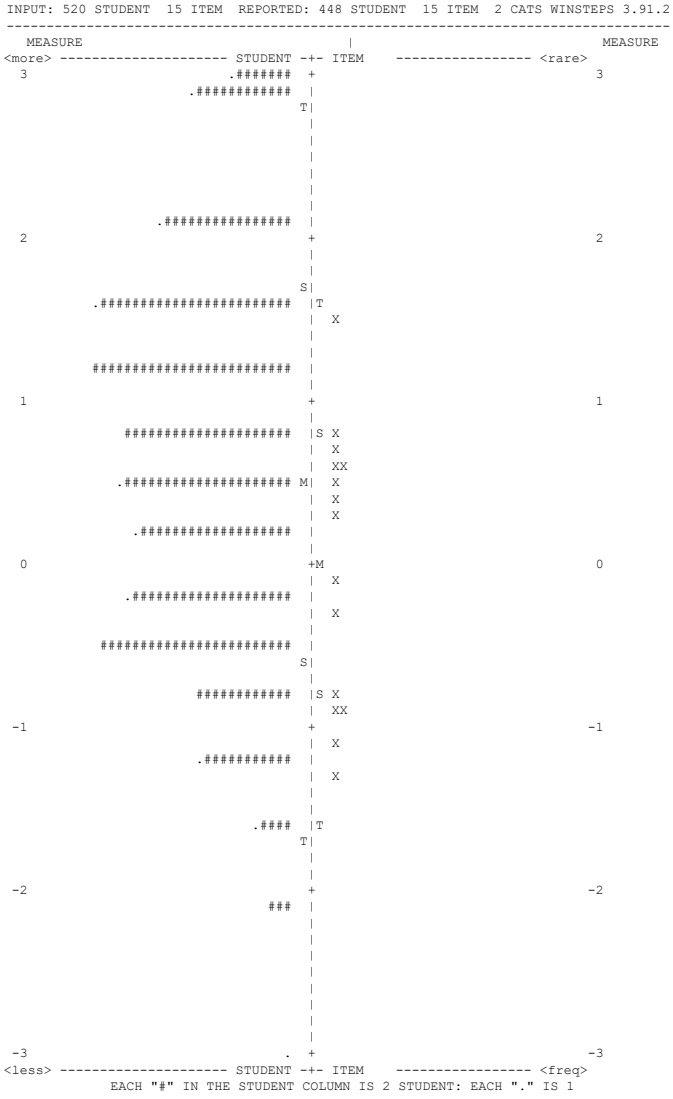

Figure 177: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2013 Lecture Test 1

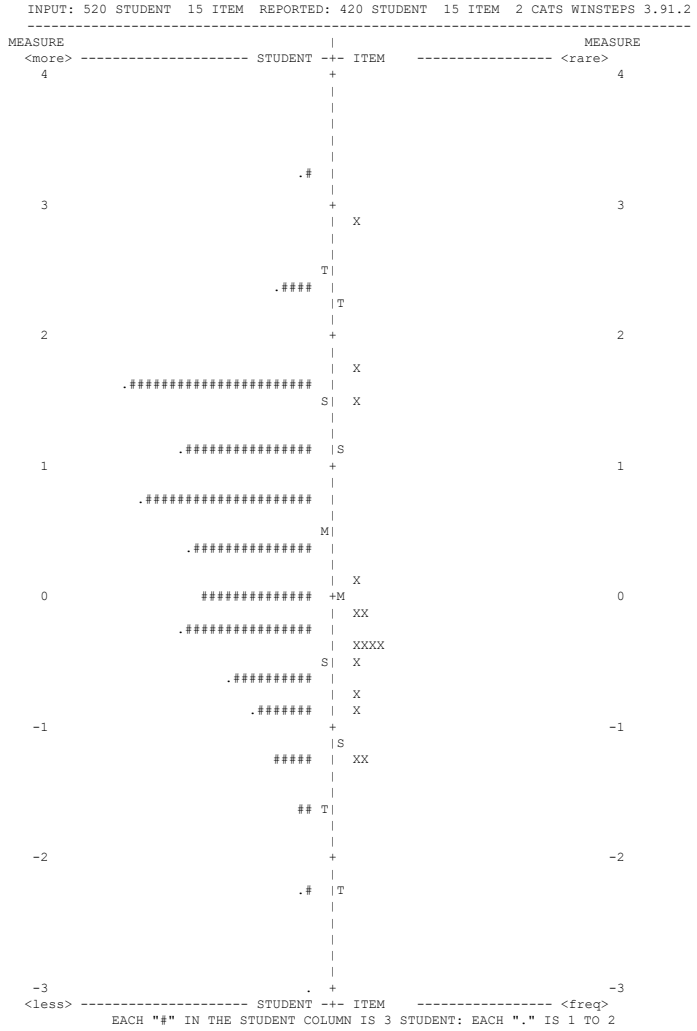

Figure 178: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2013 Lecture Test 2

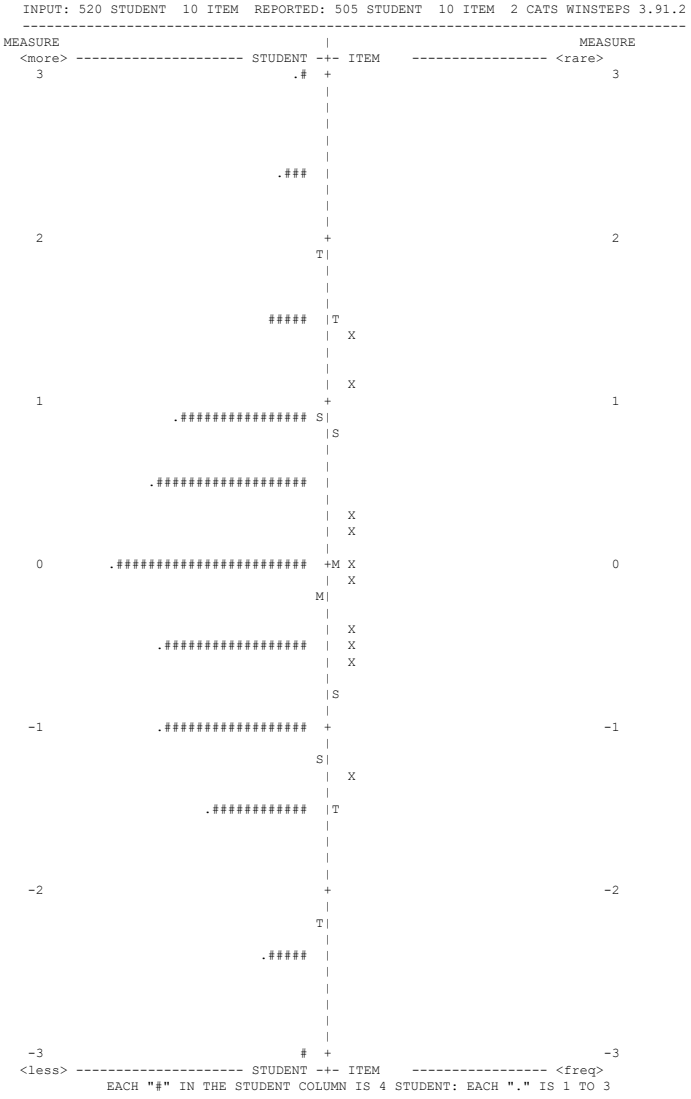

Figure 179: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2013 Exam

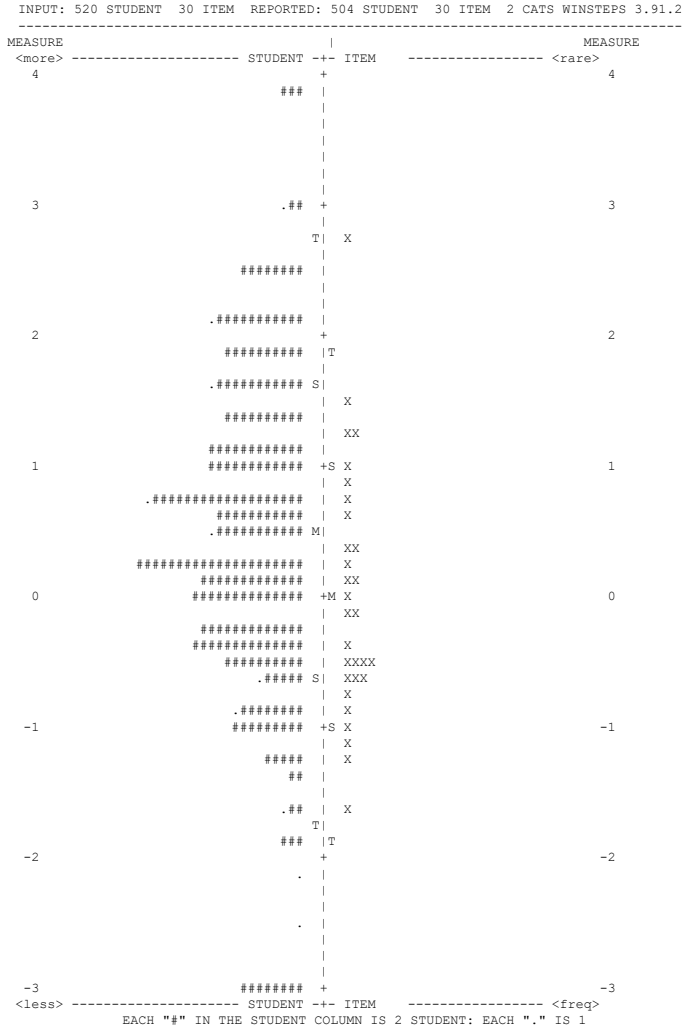

Figure 180: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2013 Redeemable Exam

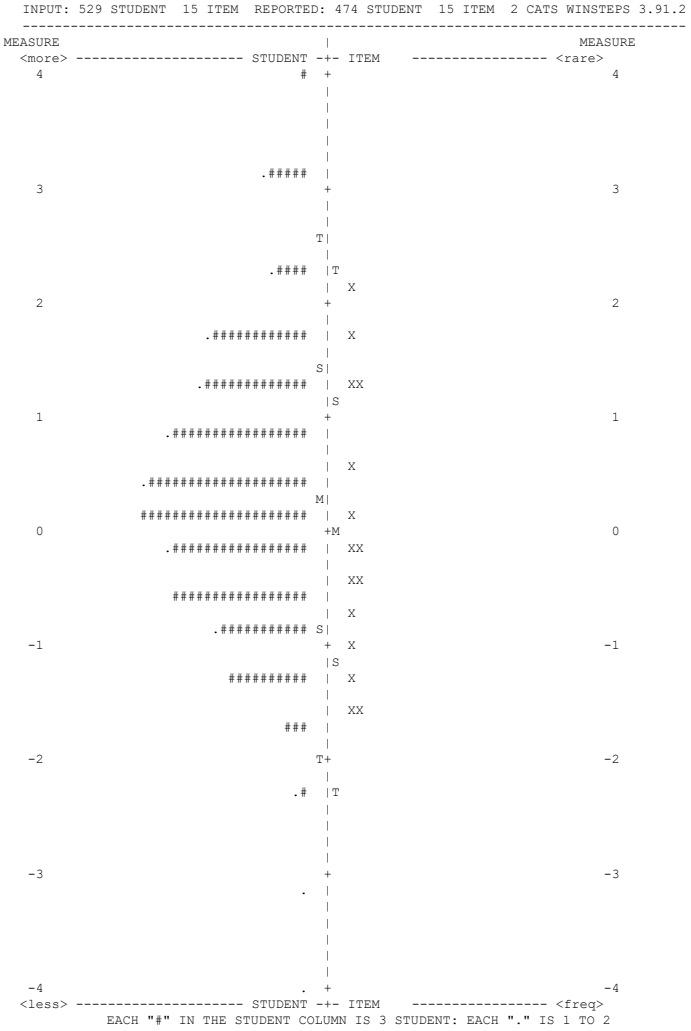

Figure 181: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2014 Lecture Test 1

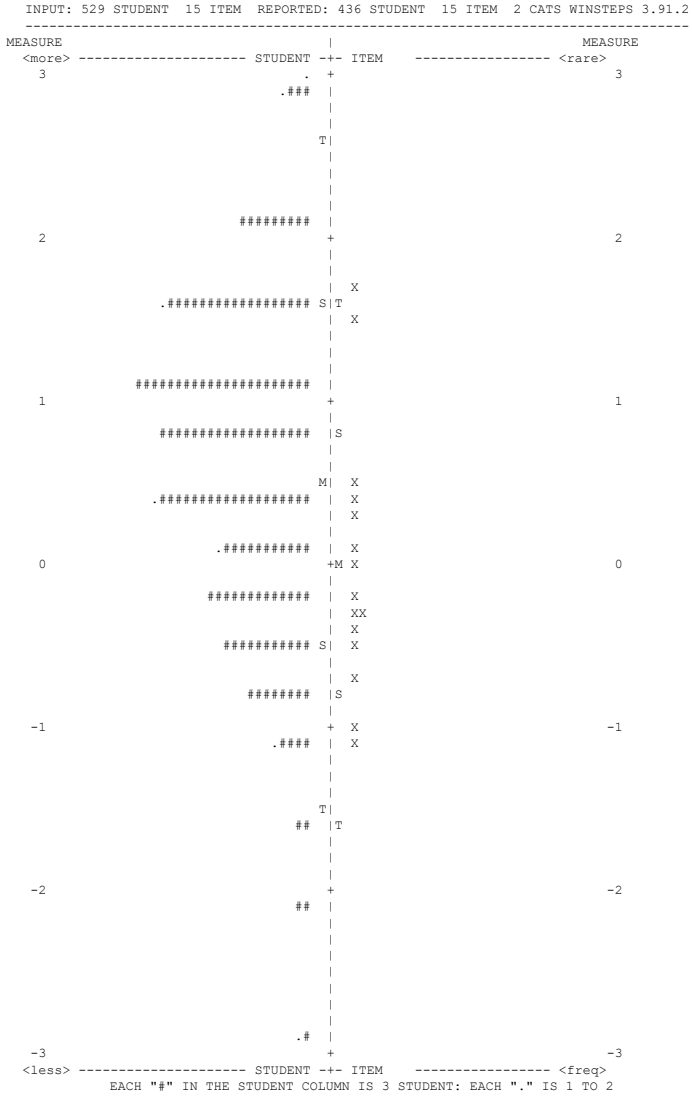

Figure 182: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2014 Lecture Test 2

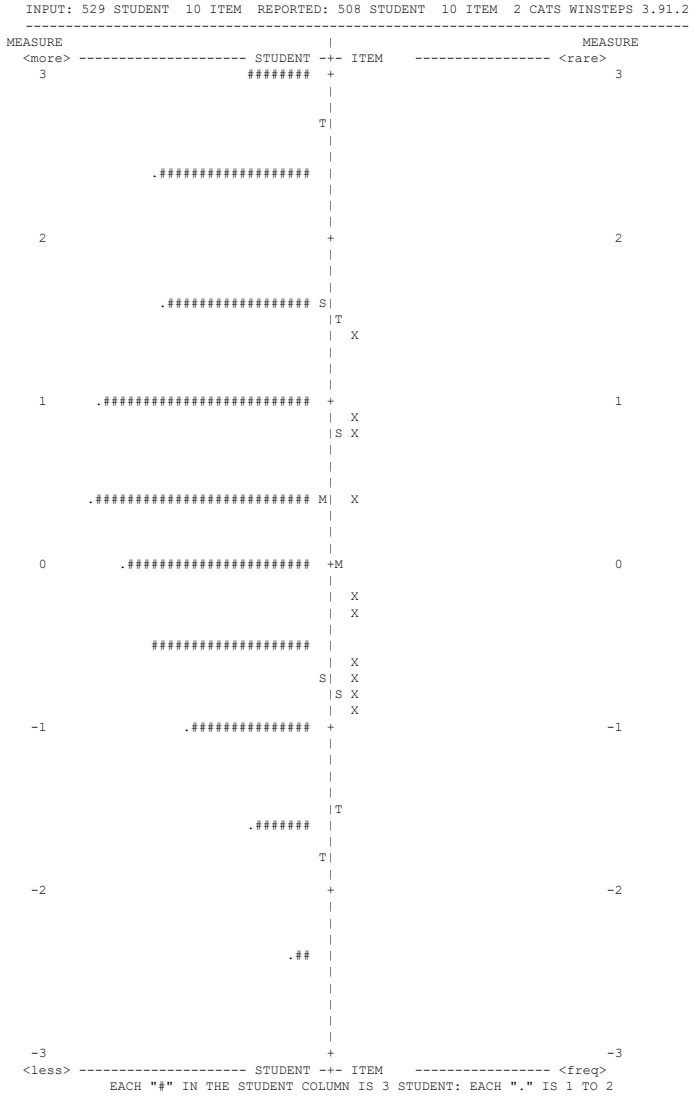

Figure 183: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2014 Exam

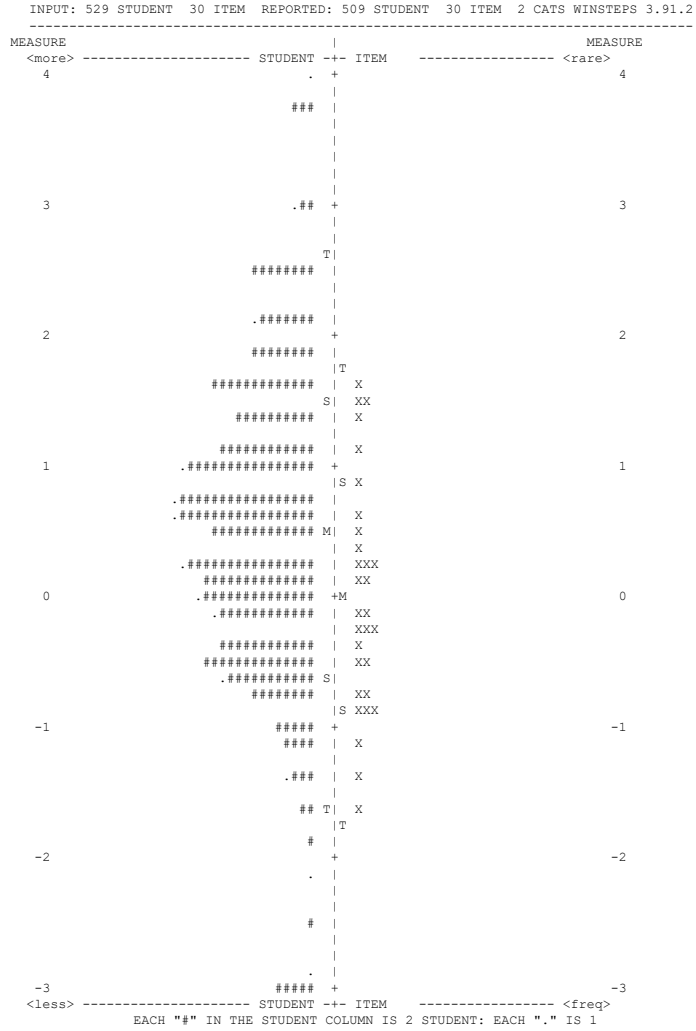

Figure 184: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2014 Redeemable Exam

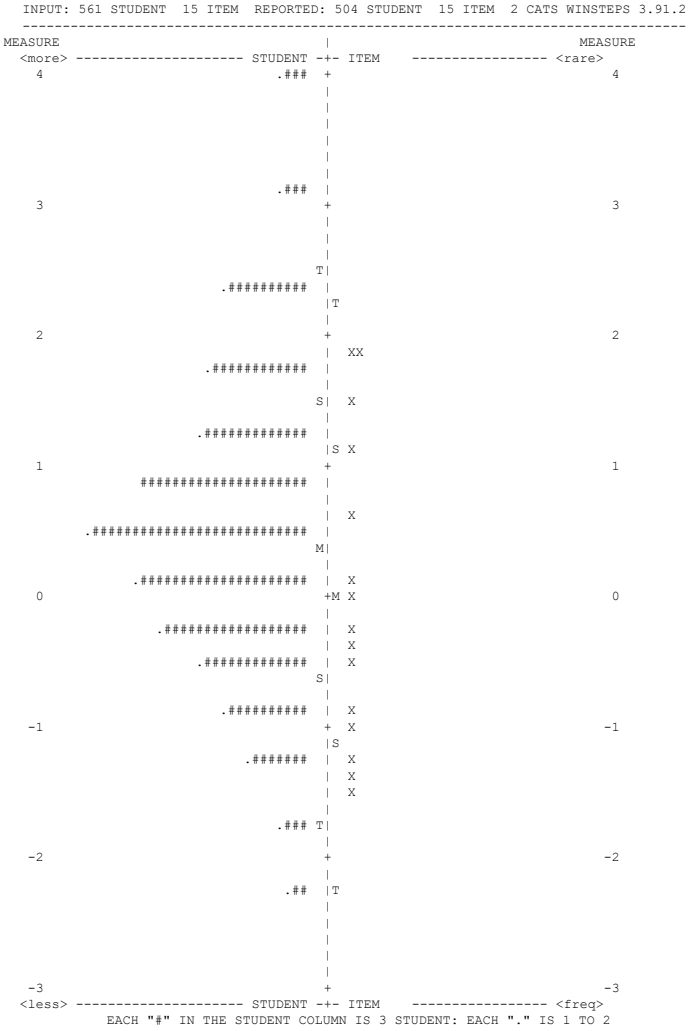

Figure 185: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2015 Lecture Test 1

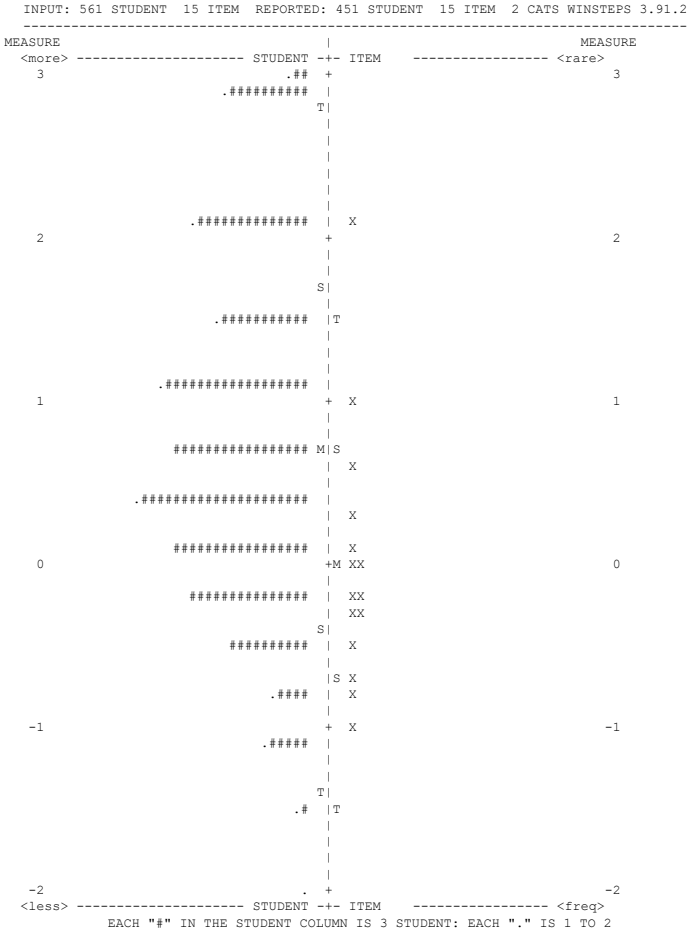

Figure 186: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2015 Lecture Test 2

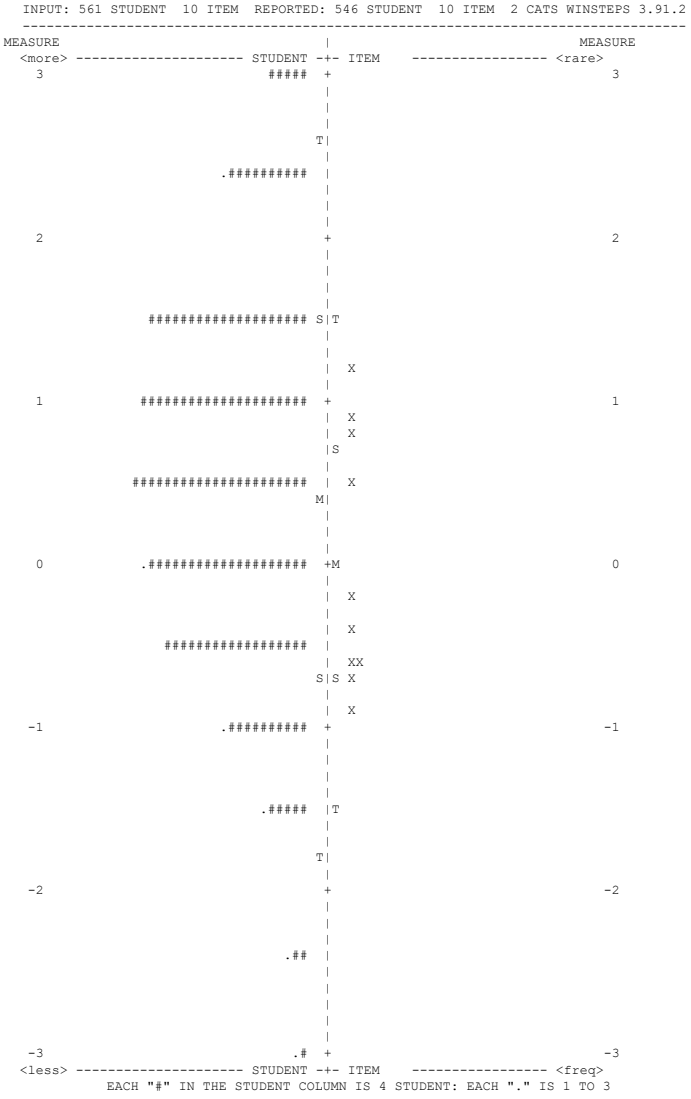

Figure 187: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2015 Exam

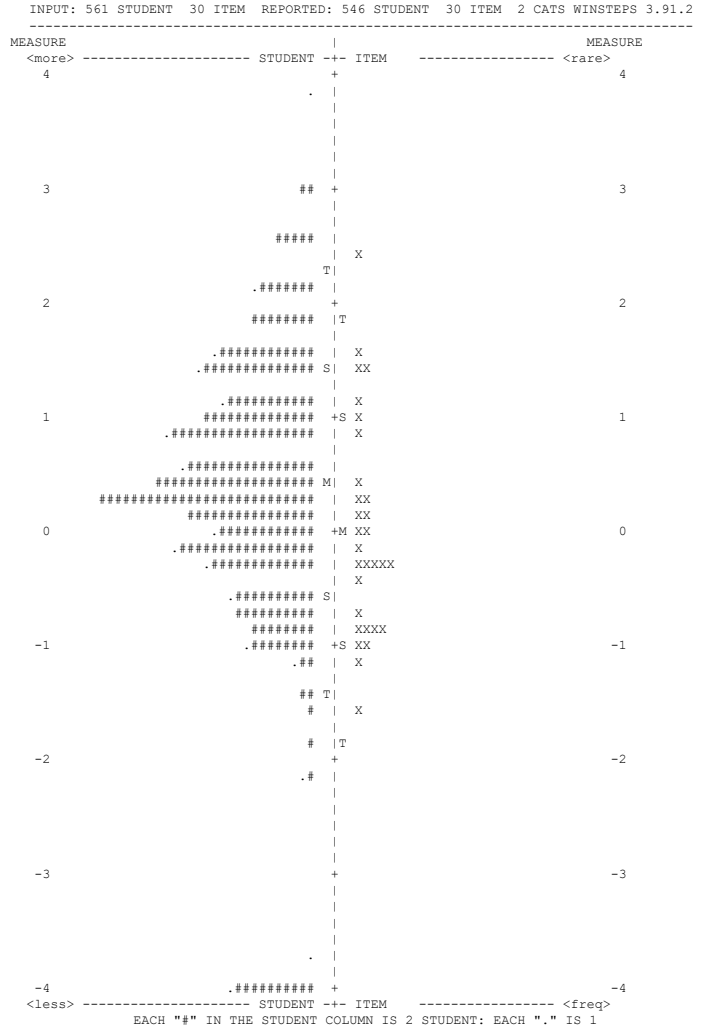

Figure 188: Wright Map of Student Ability and Item Difficulty in Chemistry IA 2015 Redeemable Exam

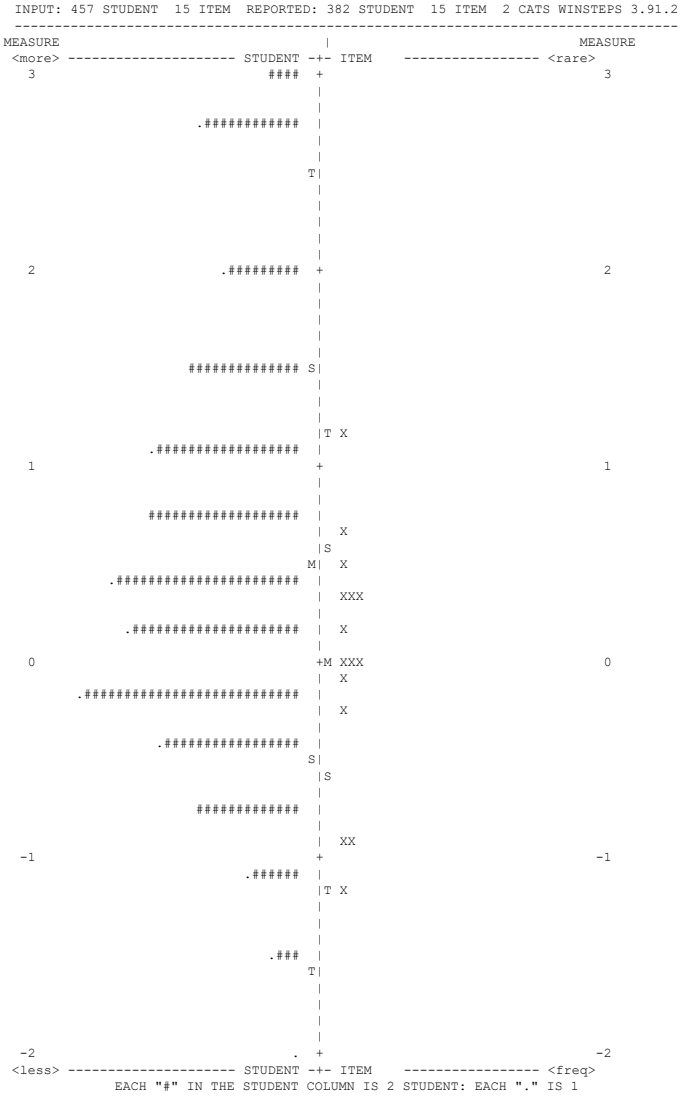

Figure 189: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2012 Lecture Test 1

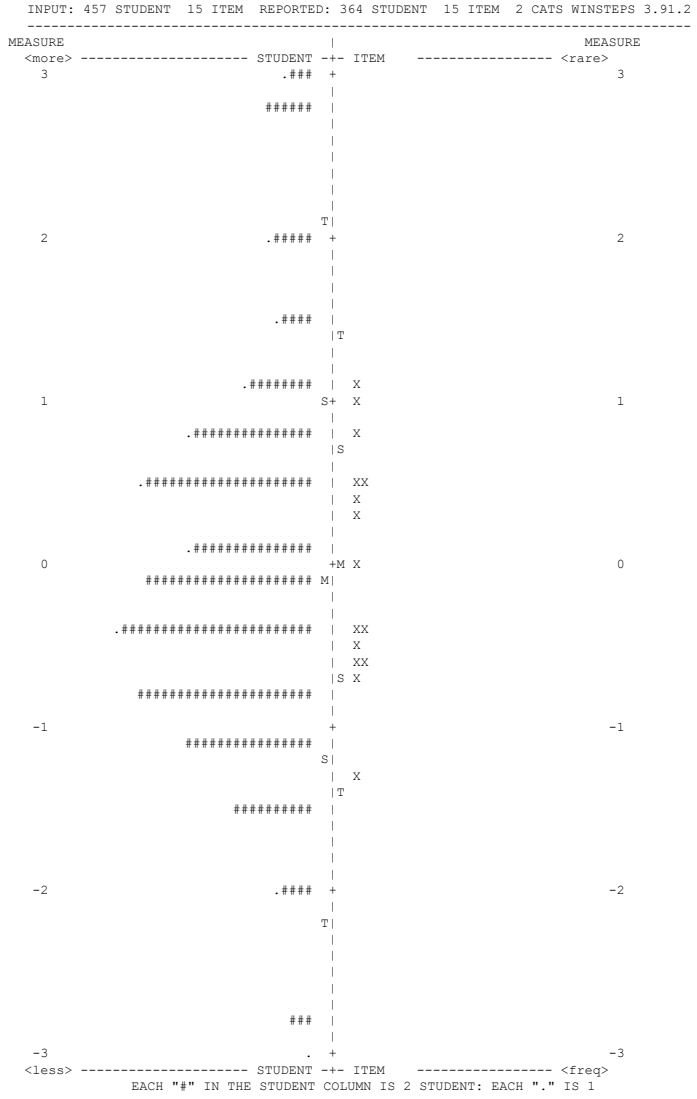

Figure 190: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2012 Lecture Test 2

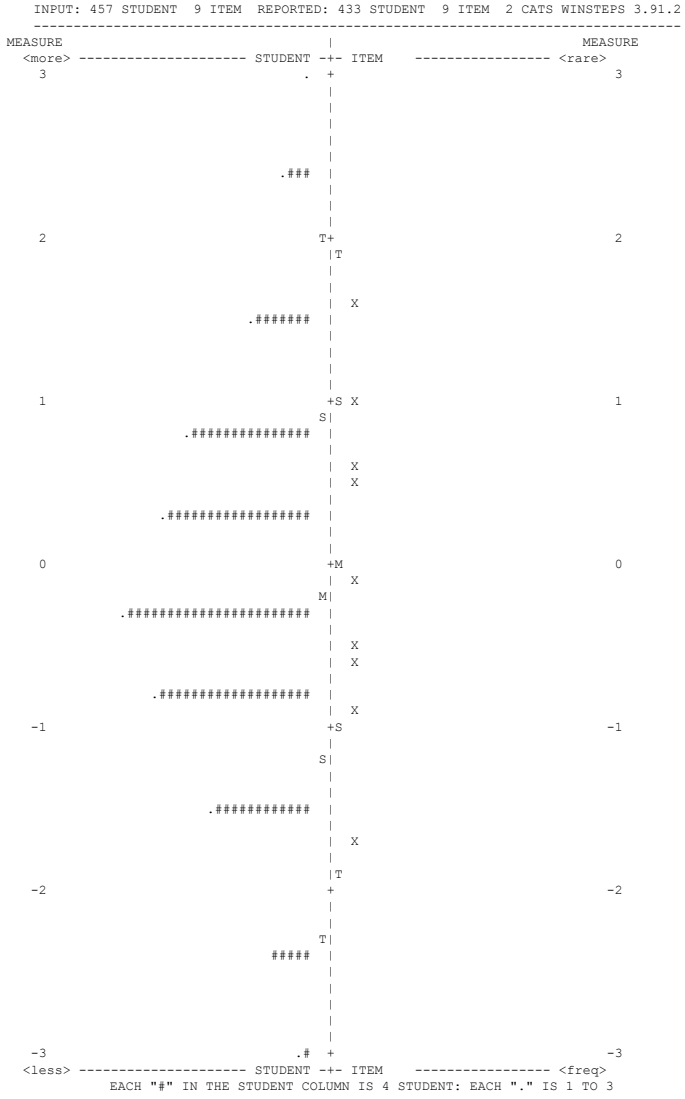

Figure 191: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2012 Exam

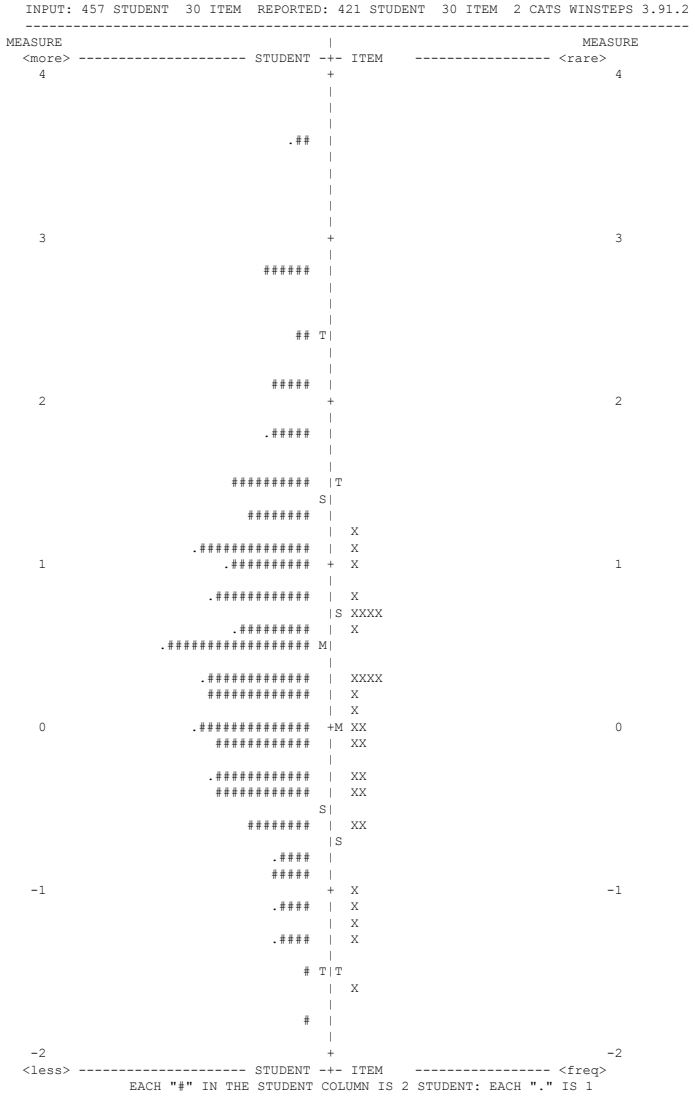

Figure 192: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2012 Redeemable Exam

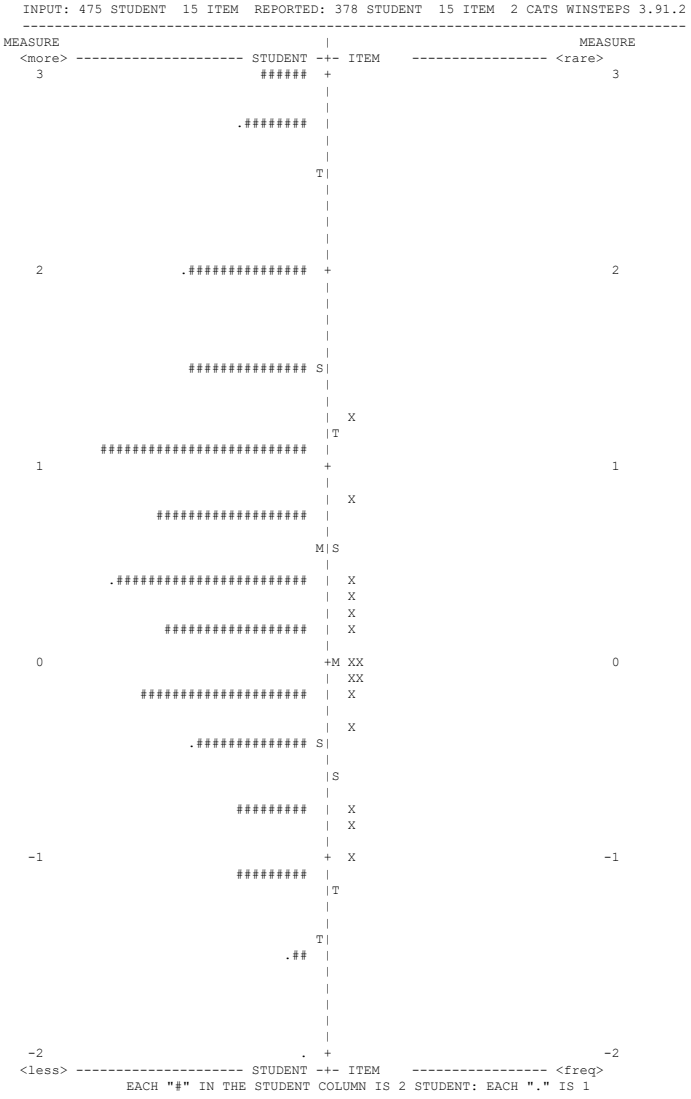

Figure 193: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2013 Lecture Test 1

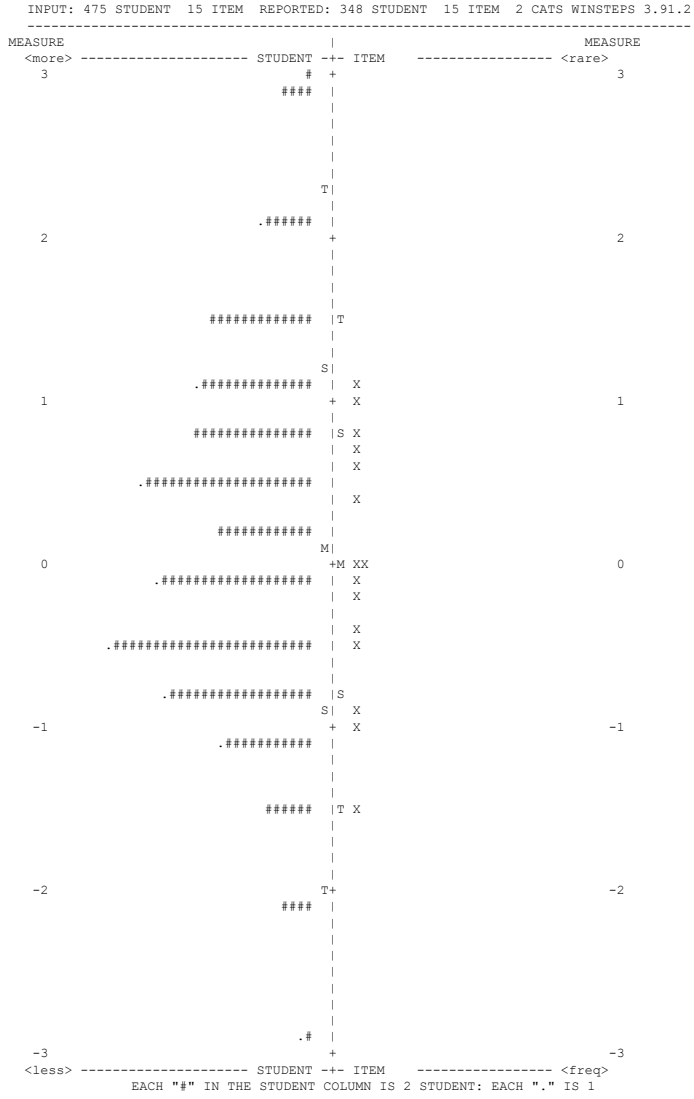

Figure 194: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2013 Lecture Test 2

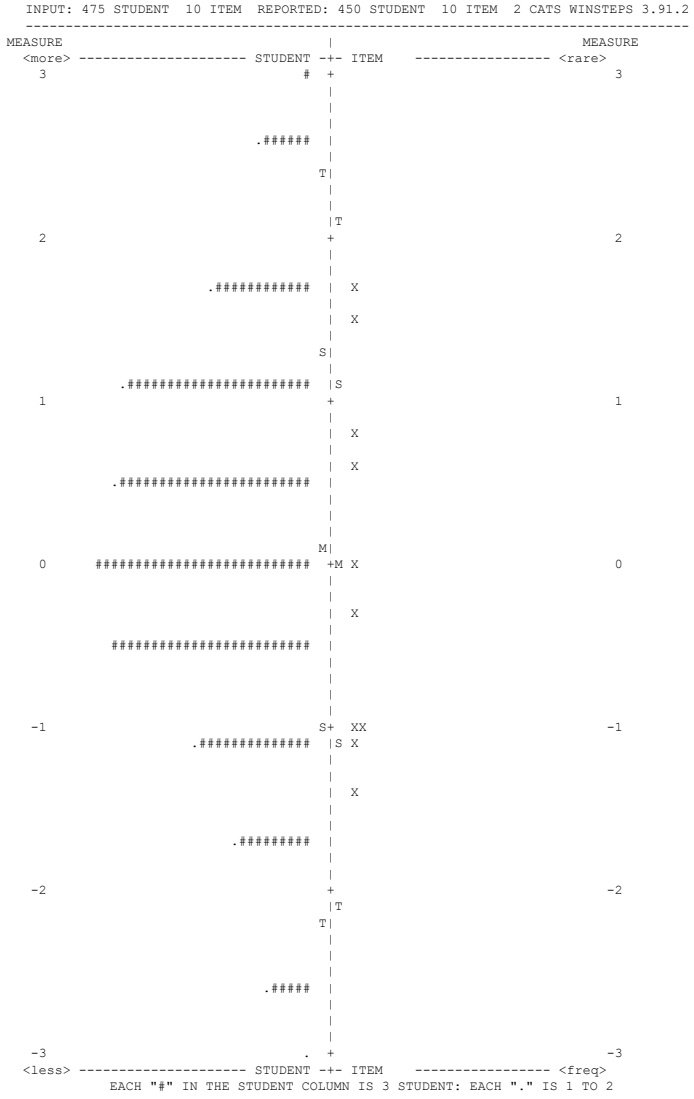

Figure 195: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2013 Exam

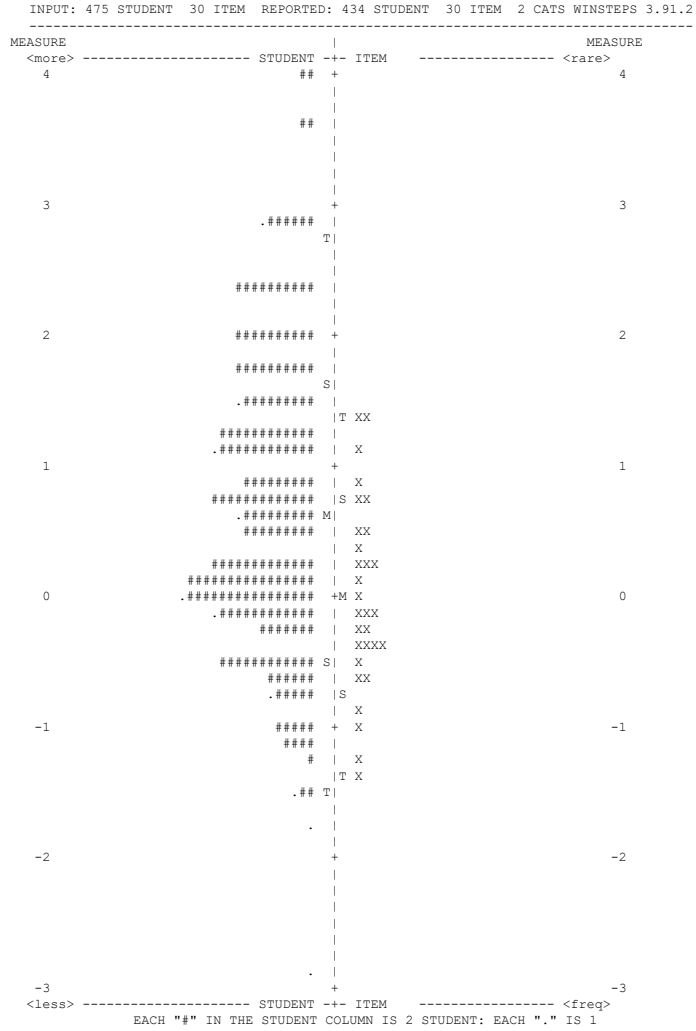

Figure 196: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2013 Redeemable Exam

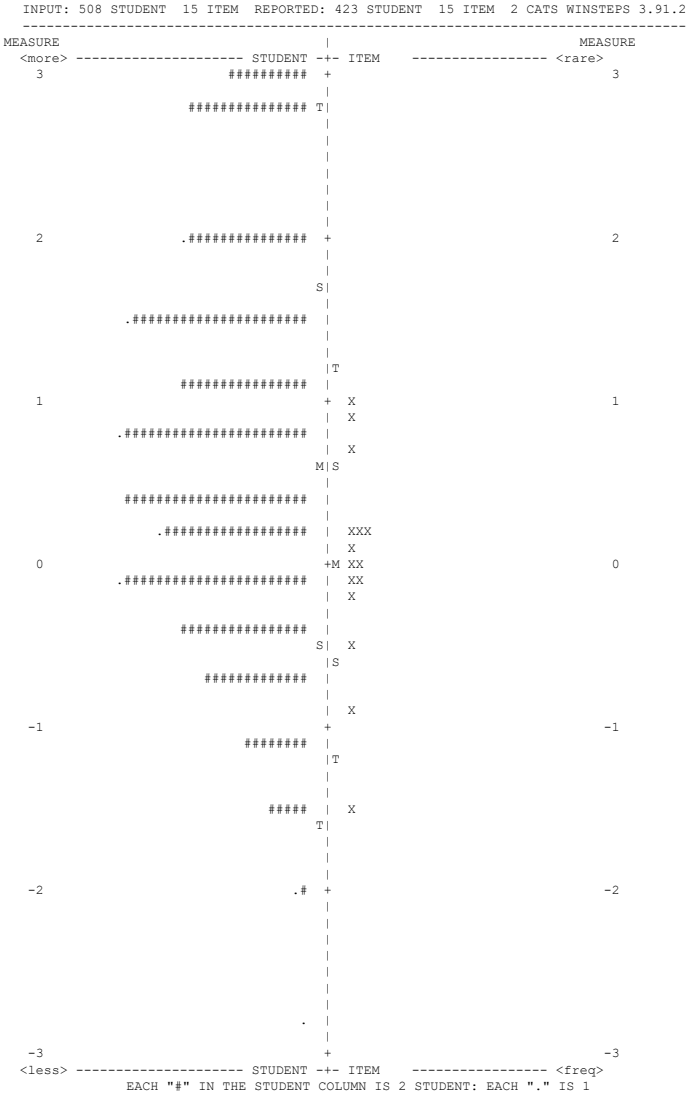

Figure 197: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2014 Lecture Test 1

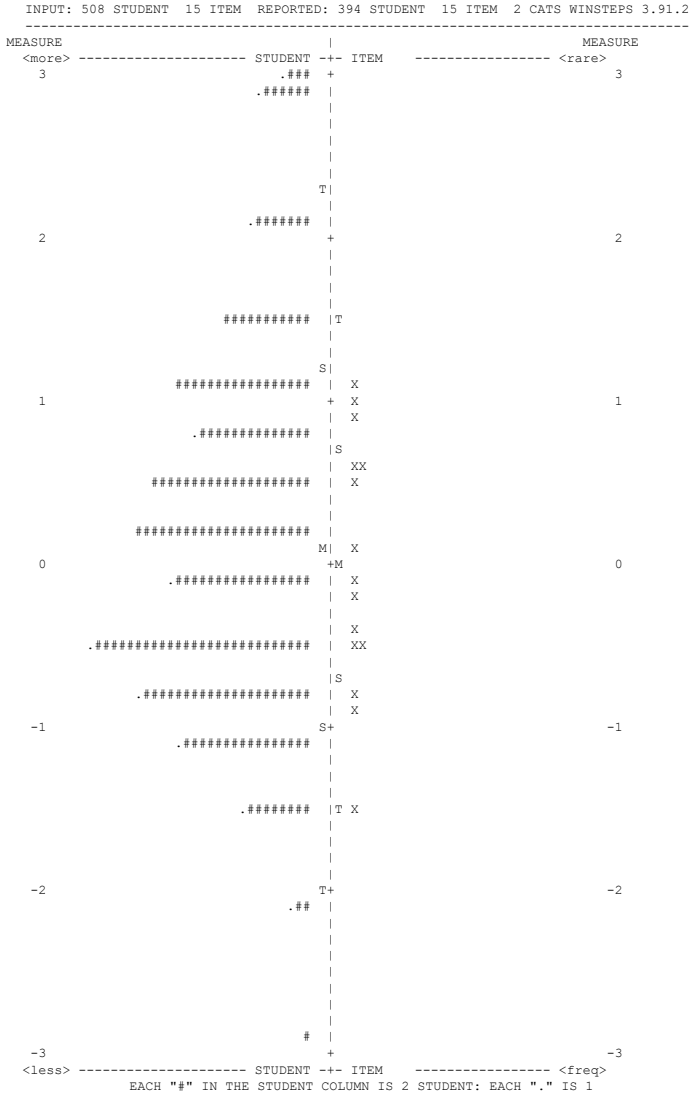

Figure 198: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2014 Lecture Test 2

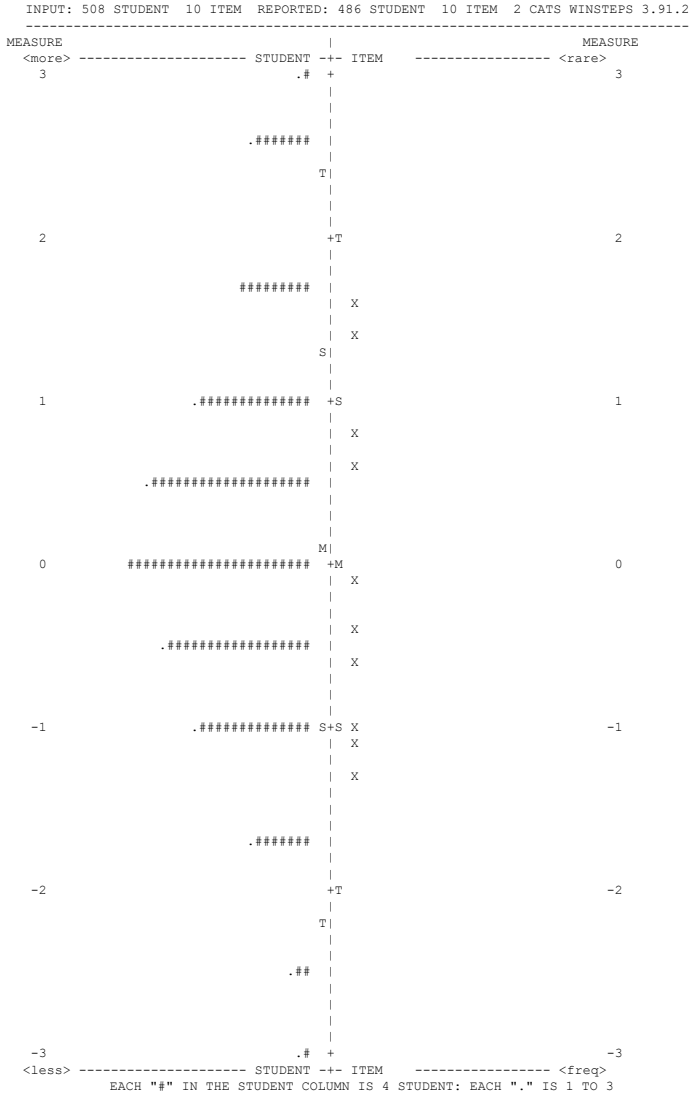

Figure 199: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2014 Exam

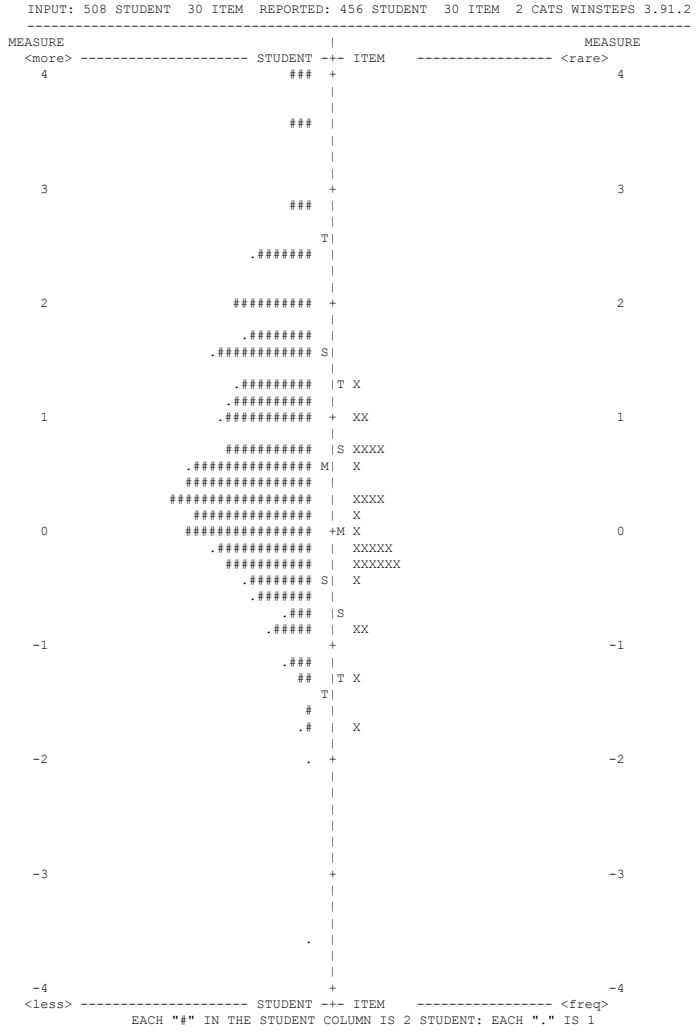

Figure 200: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2014 Redeemable Exam

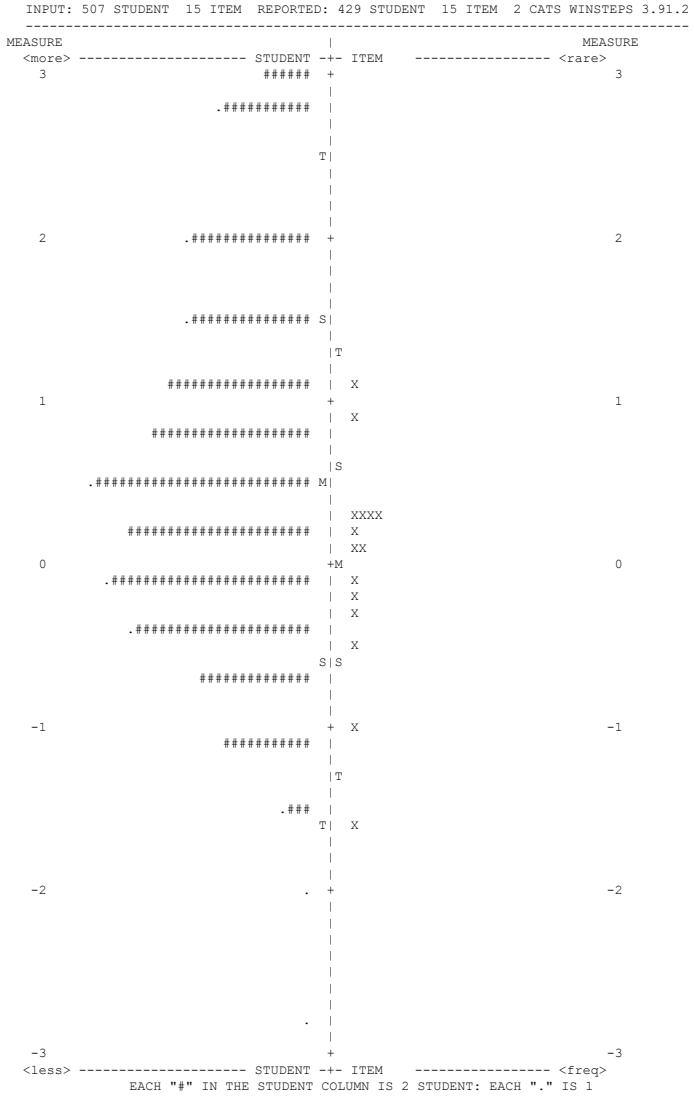

Figure 201: Wright Map of Student Ability and Item Difficulty in Chemistry
IB 2015 Lecture Test 1

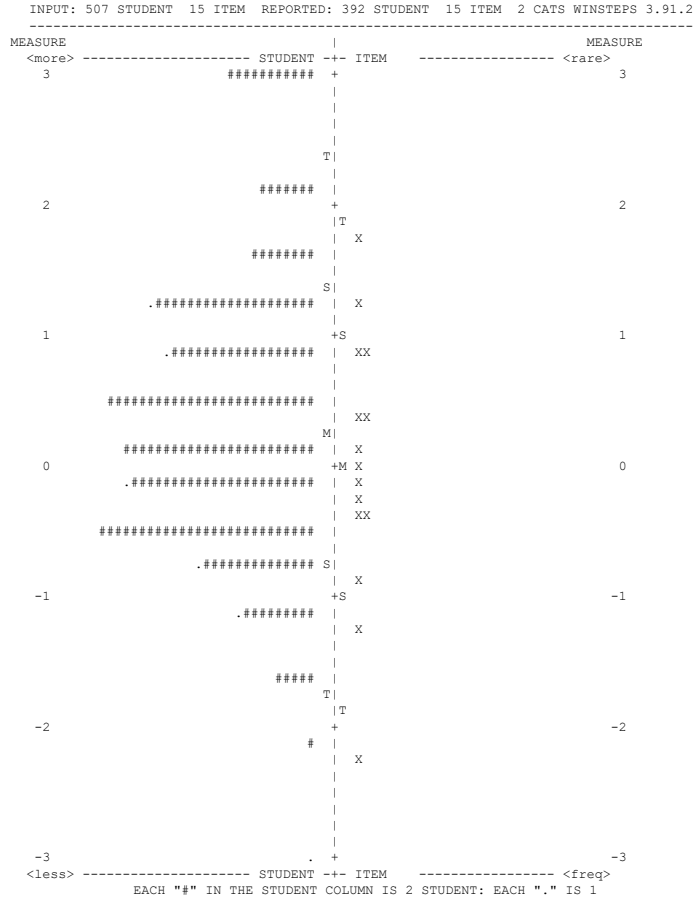

Figure 202: Wright Map of Student Ability and Item Difficulty in Chemistry
IB 2015 Lecture Test 2

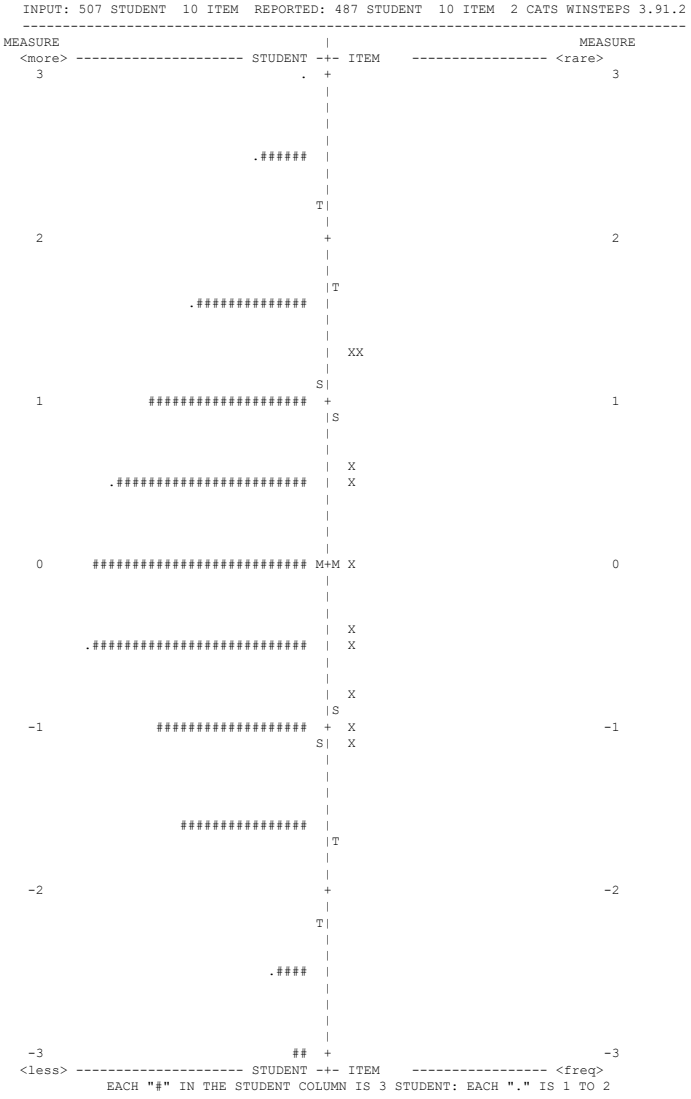

Figure 203: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2015 Exam

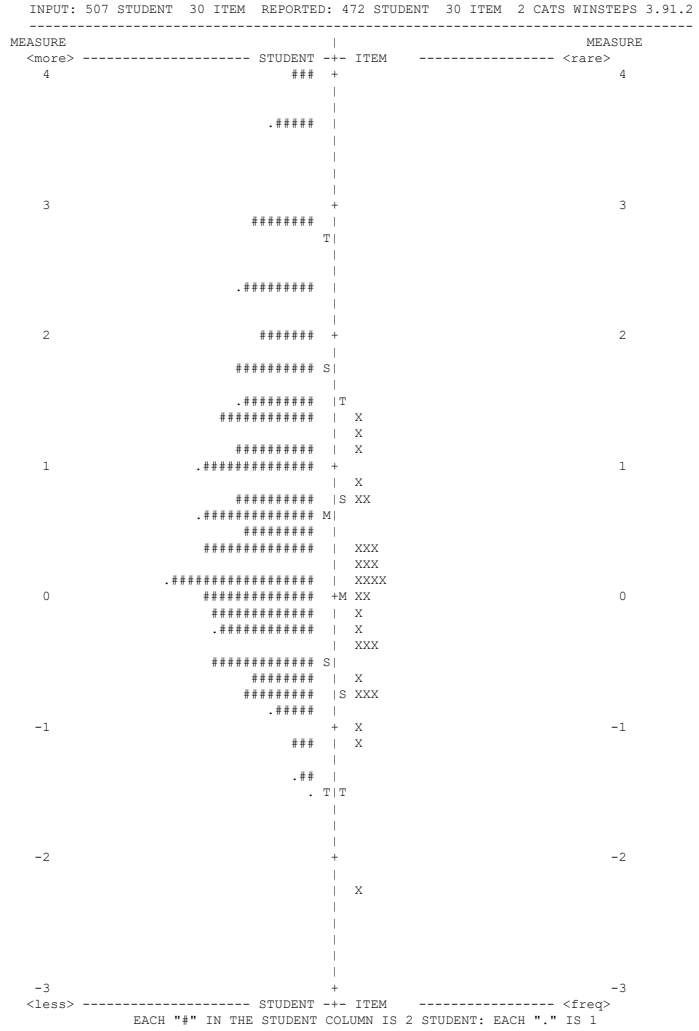

Figure 204: Wright Map of Student Ability and Item Difficulty in Chemistry IB 2015 Redeemable Exam

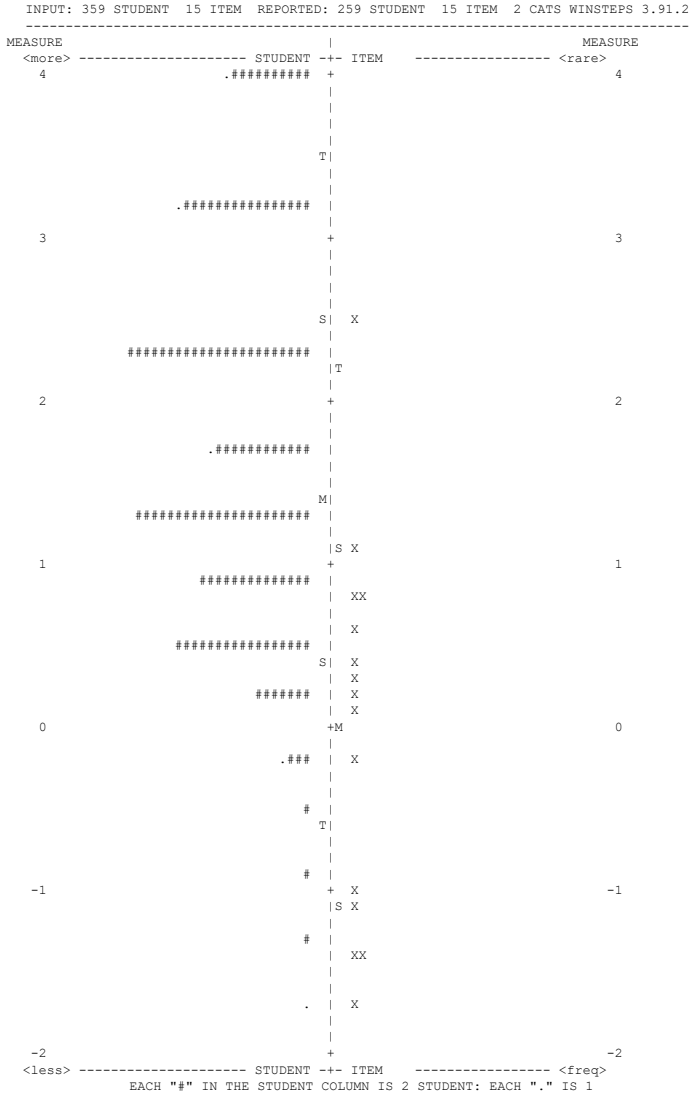

Figure 205: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2012 Lecture Test 1

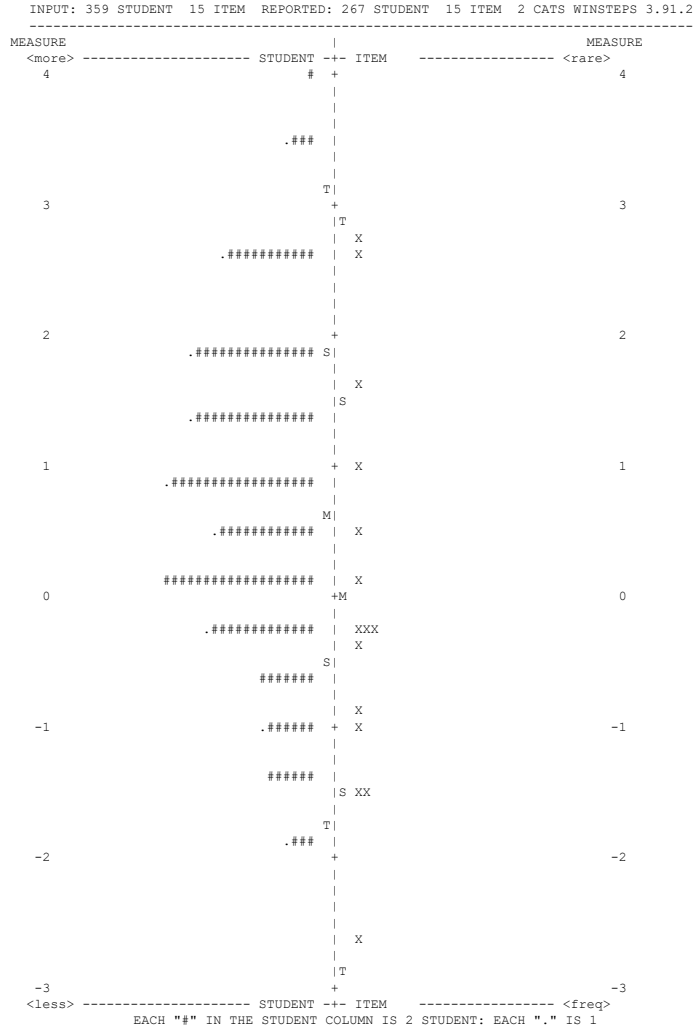

Figure 206: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2012 Lecture Test 2

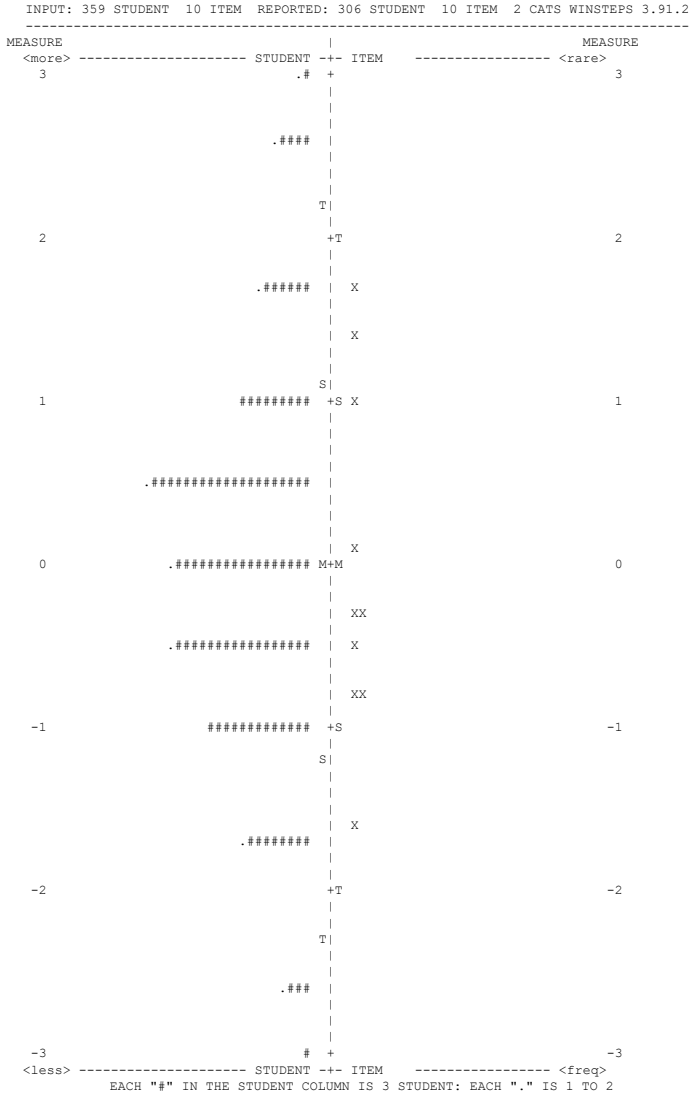

Figure 207: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2012 Exam

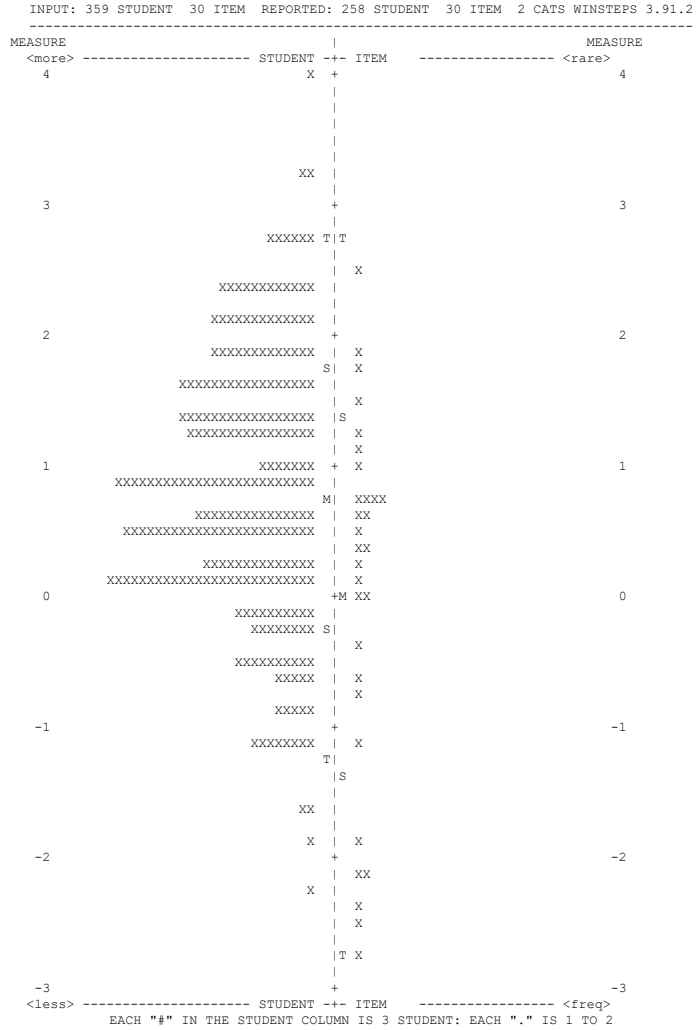

Figure 208: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2012 Redeemable Exam

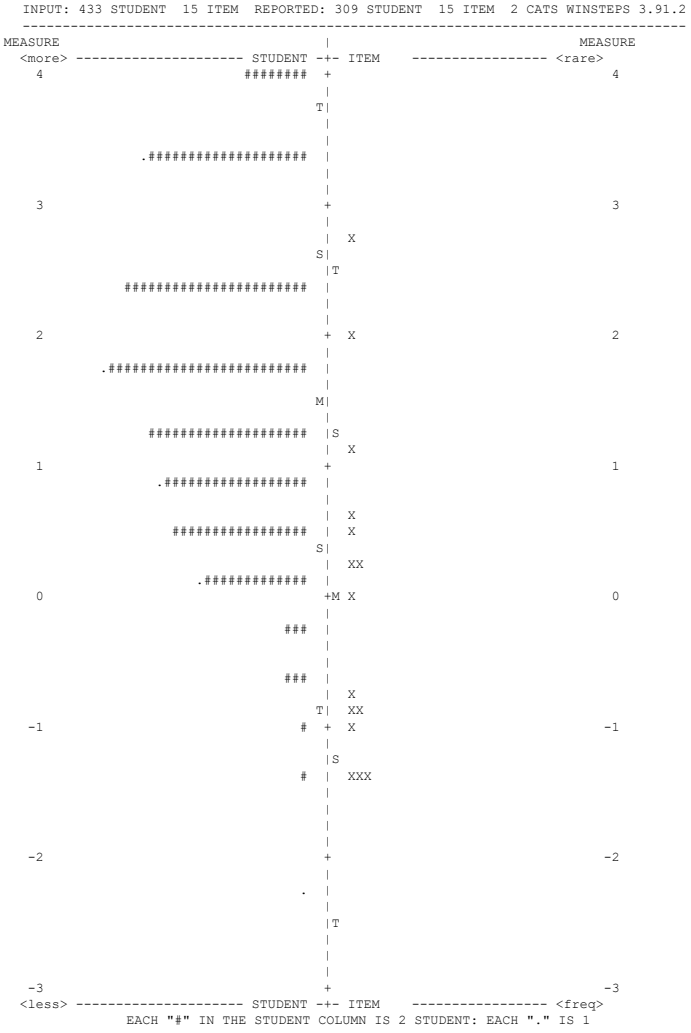

Figure 209: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2013 Lecture Test 1

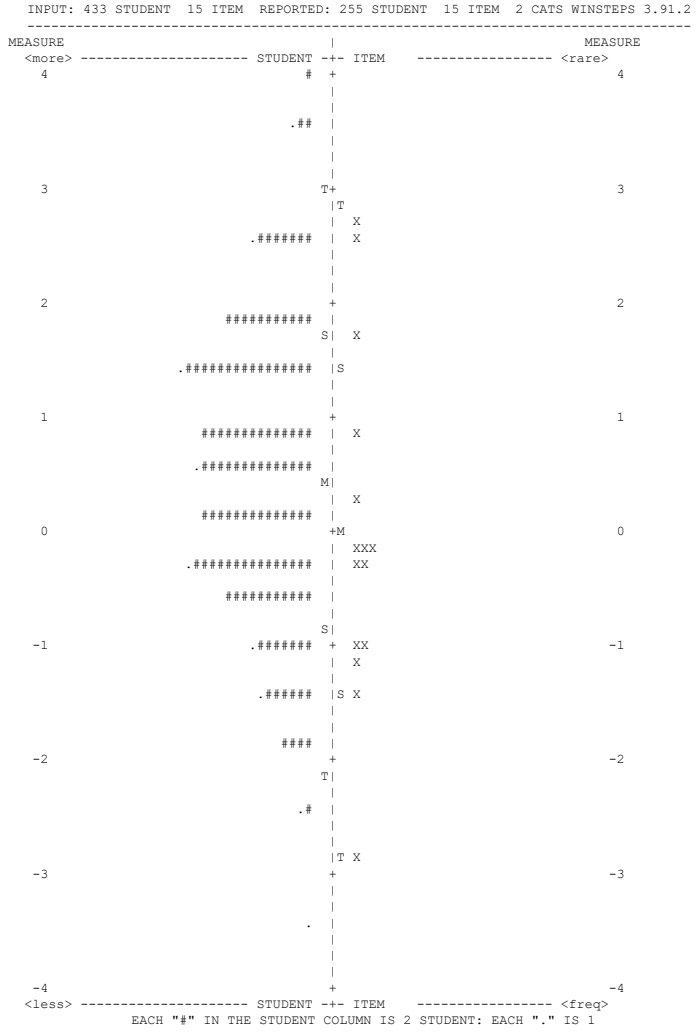

Figure 210: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2013 Lecture Test 2

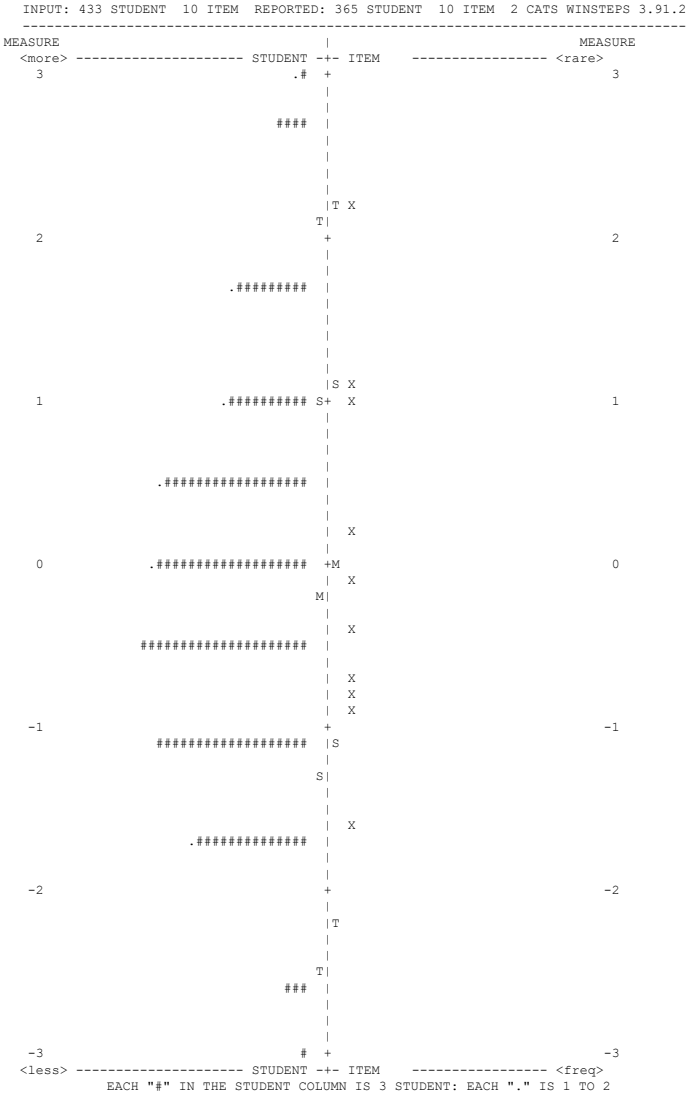

Figure 211: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2013 Exam

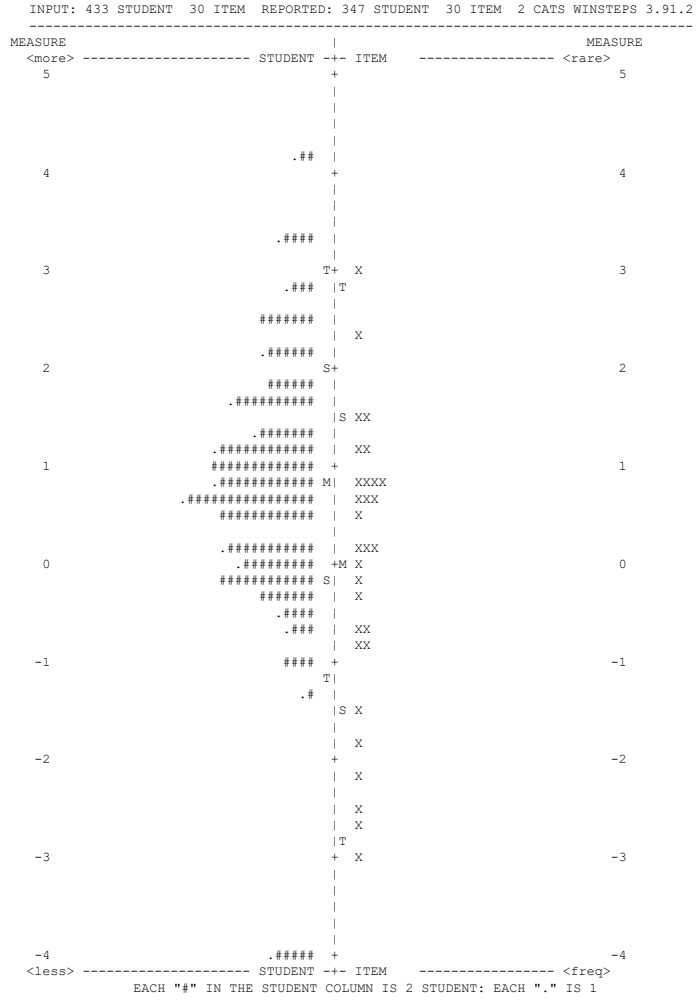

Figure 212: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2013 Redeemable Exam

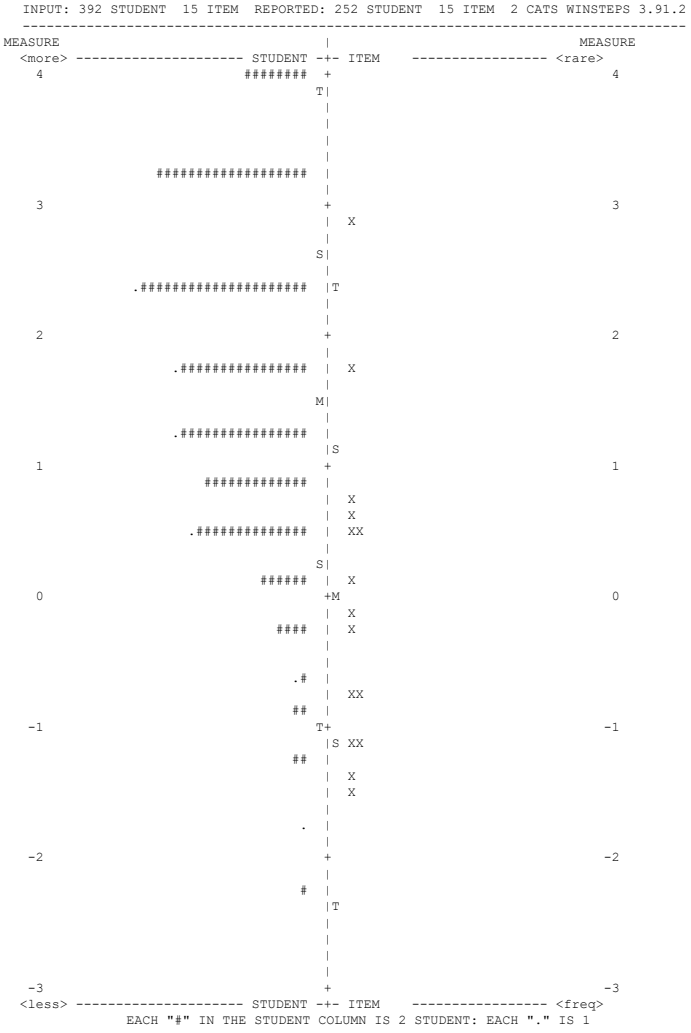

Figure 213: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2014 Lecture Test 1

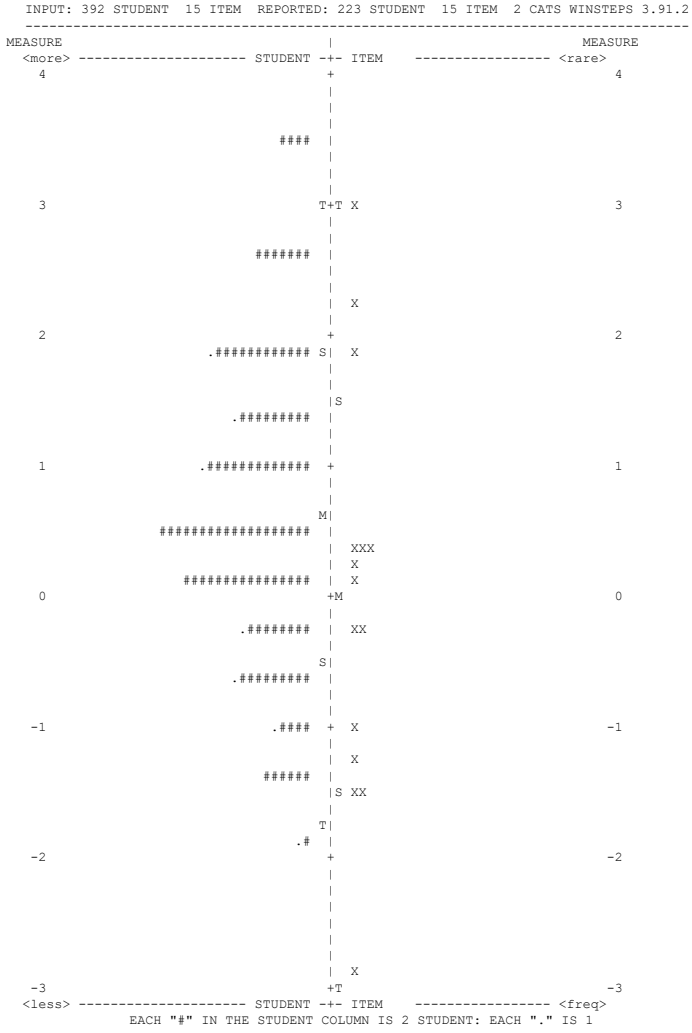

Figure 214: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2014 Lecture Test 2

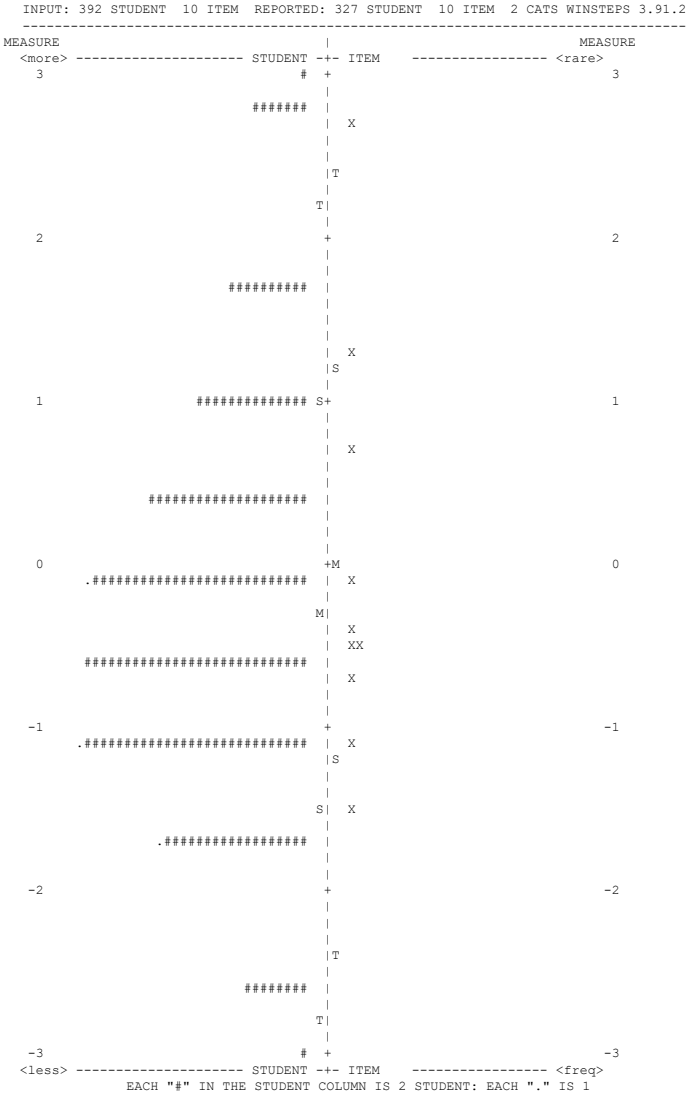

Figure 215: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2014 Exam

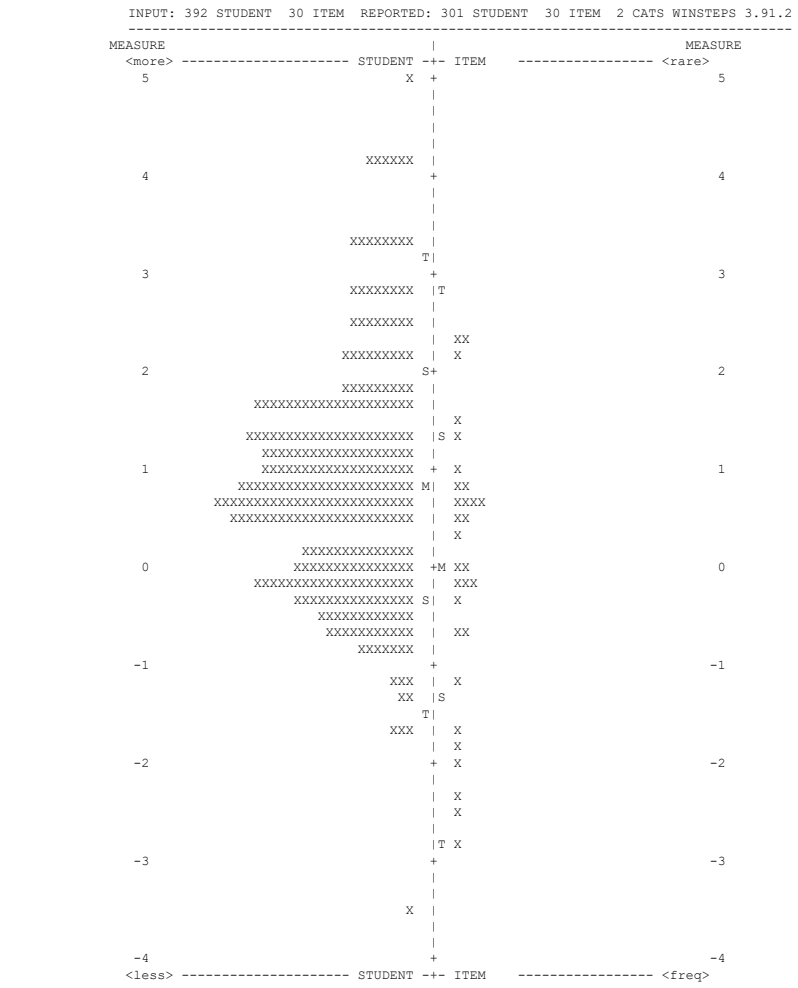

Figure 216: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2014 Redeemable Exam

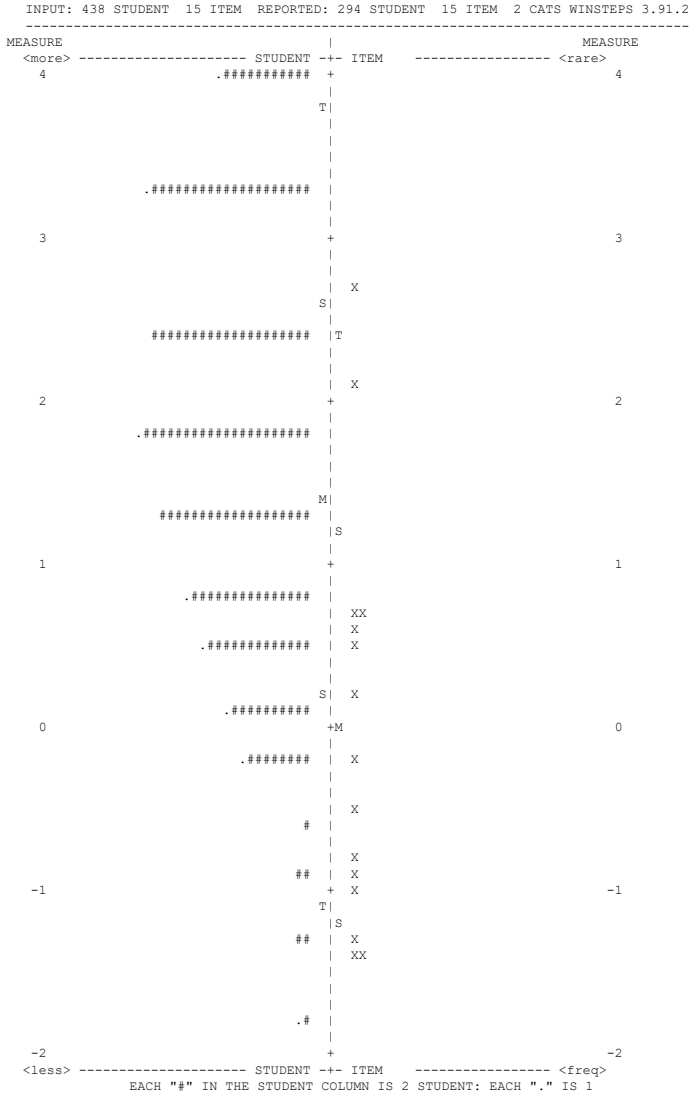

Figure 217: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2015 Lecture Test 1

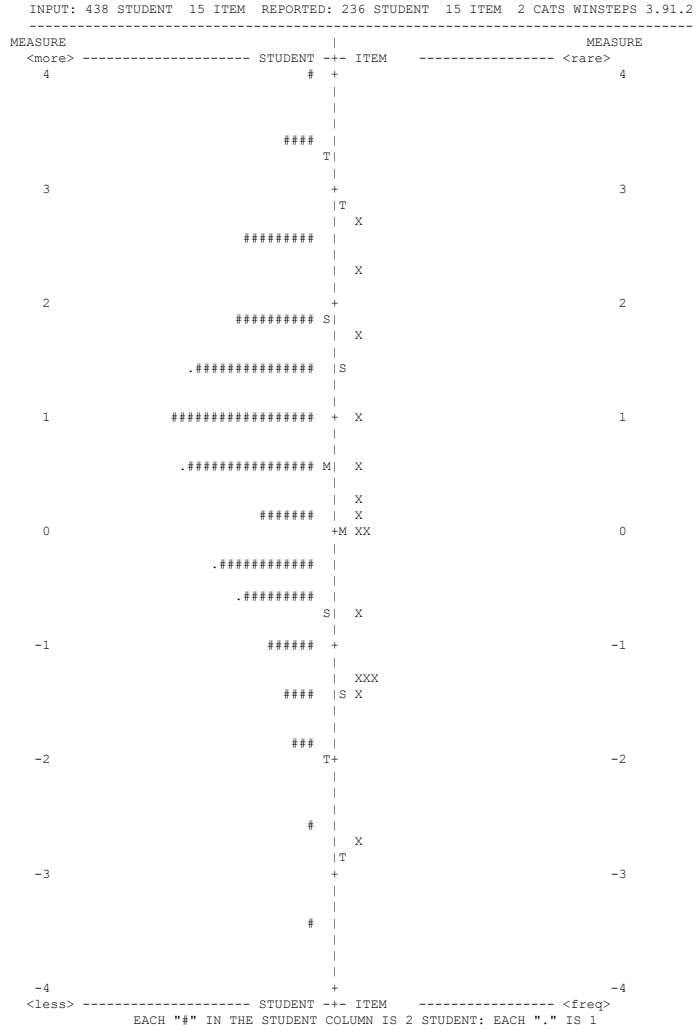

Figure 218: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2015 Lecture Test 2

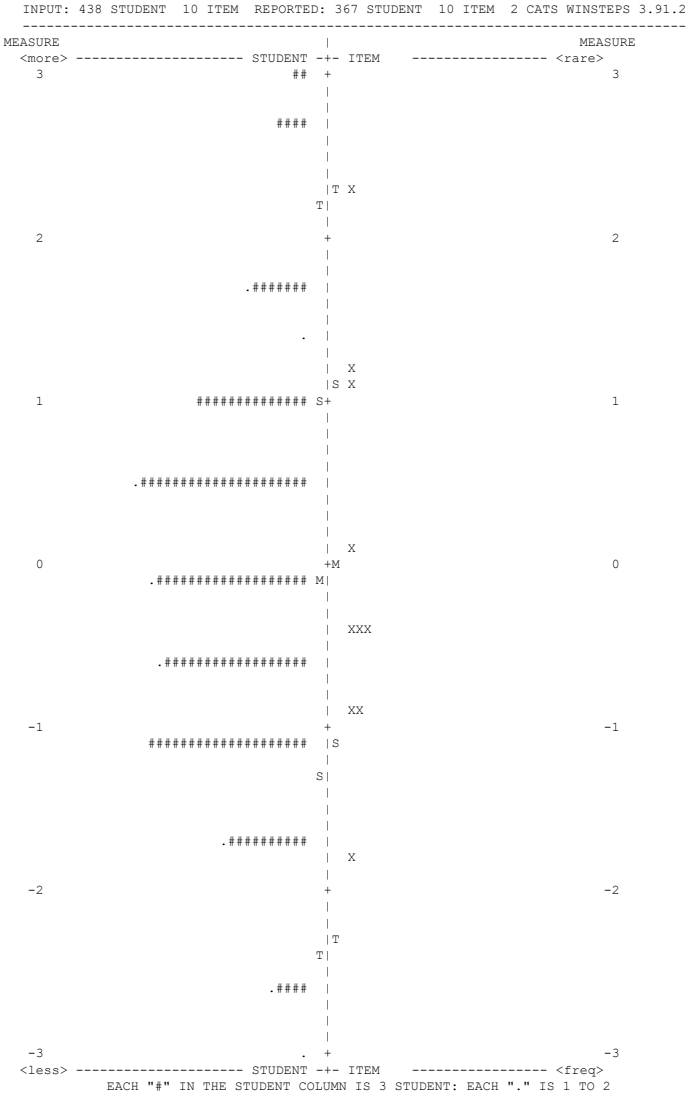

Figure 219: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2015 Exam

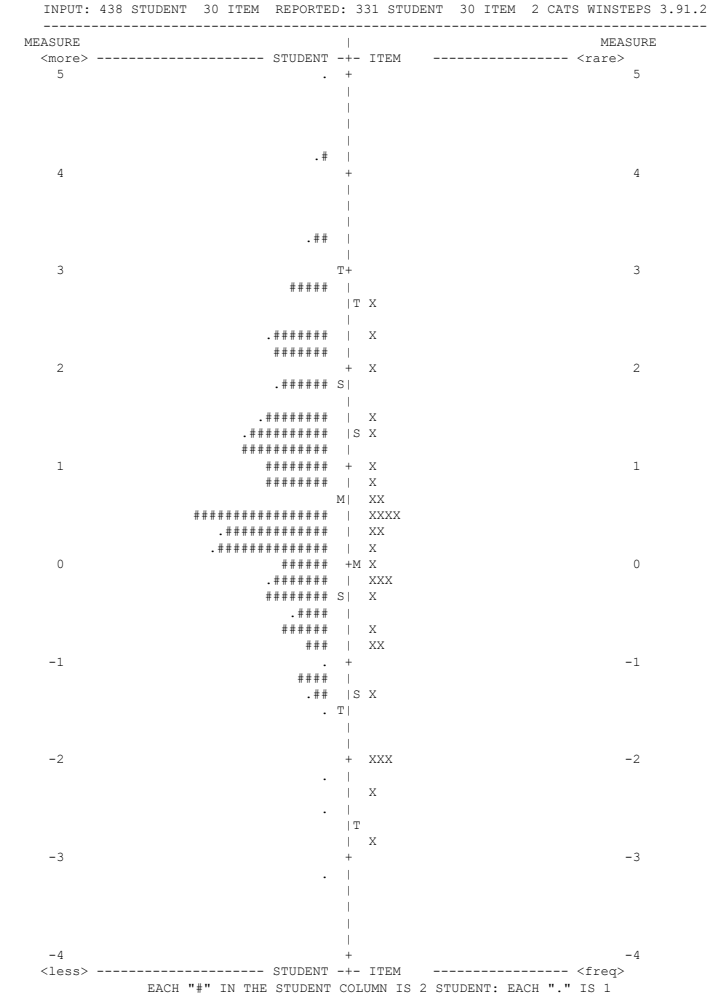

Figure 220: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IA 2015 Redeemable Exam

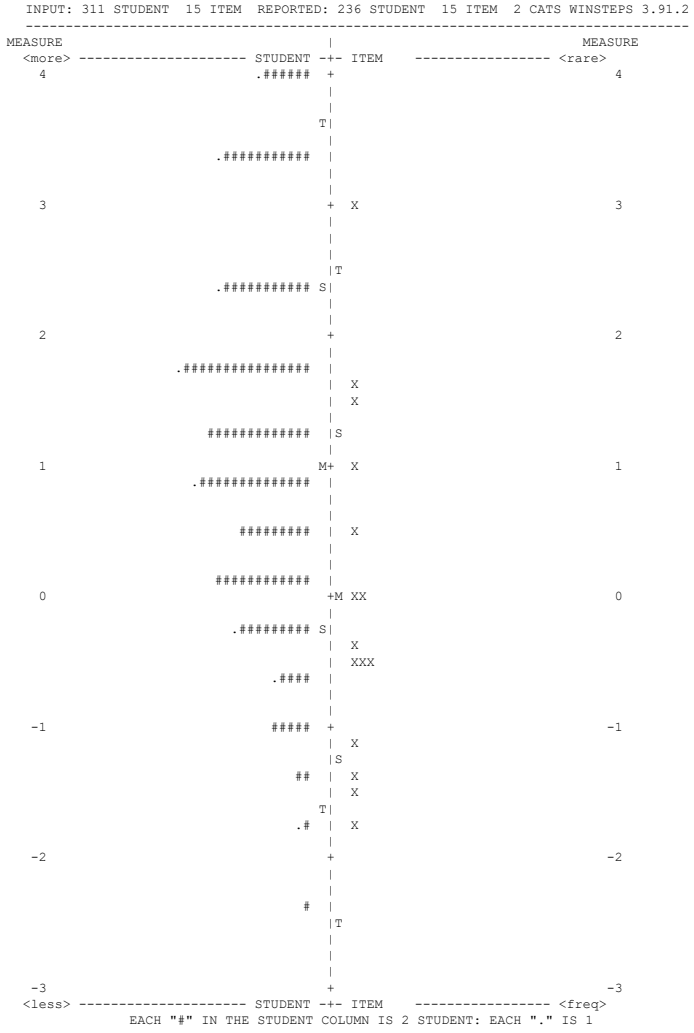

Figure 221: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2012 Lecture Test 1

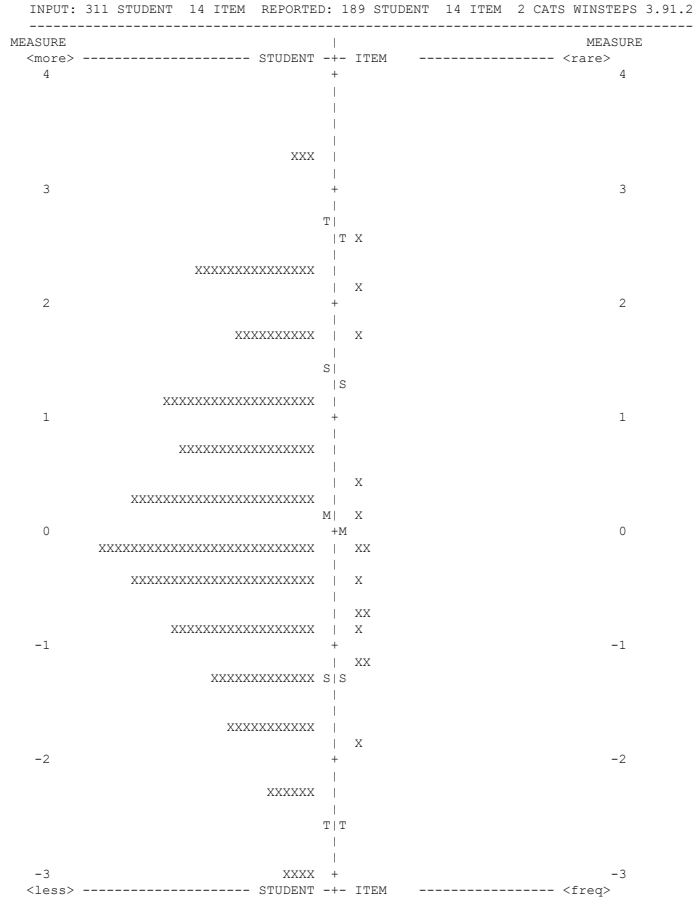

Figure 222: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2012 Lecture Test 2

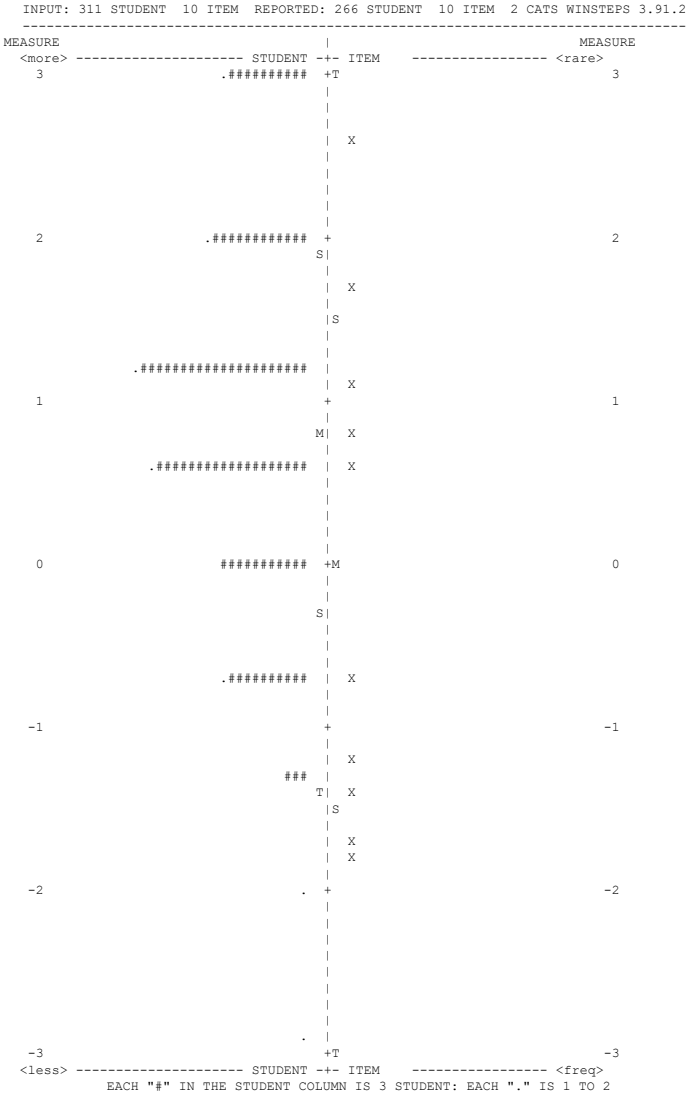

Figure 223: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2012 Exam

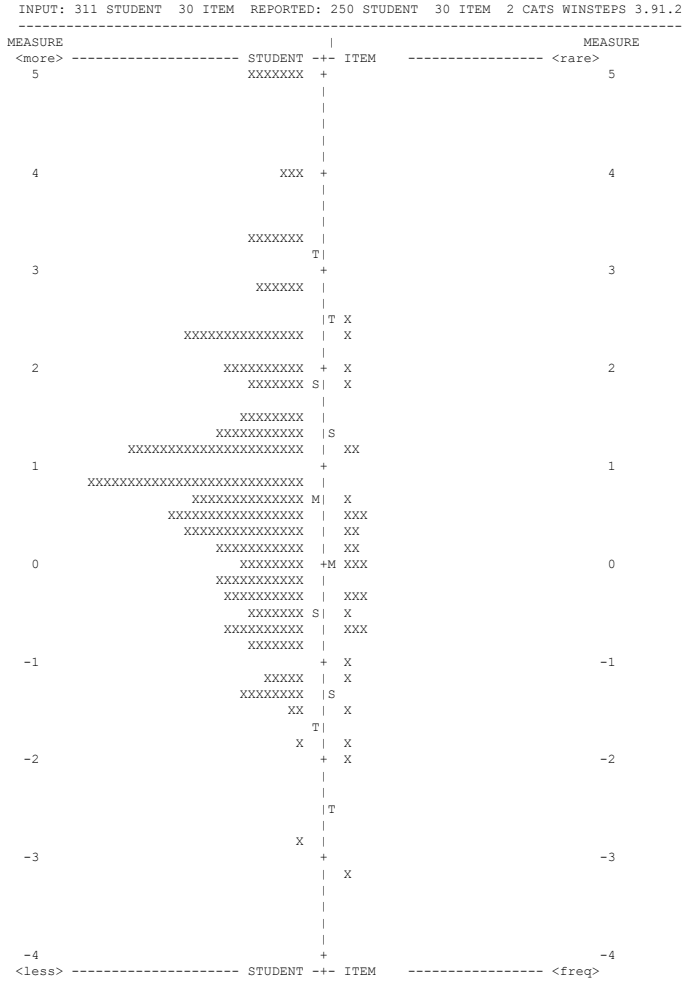

Figure 224: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2012 Redeemable Exam

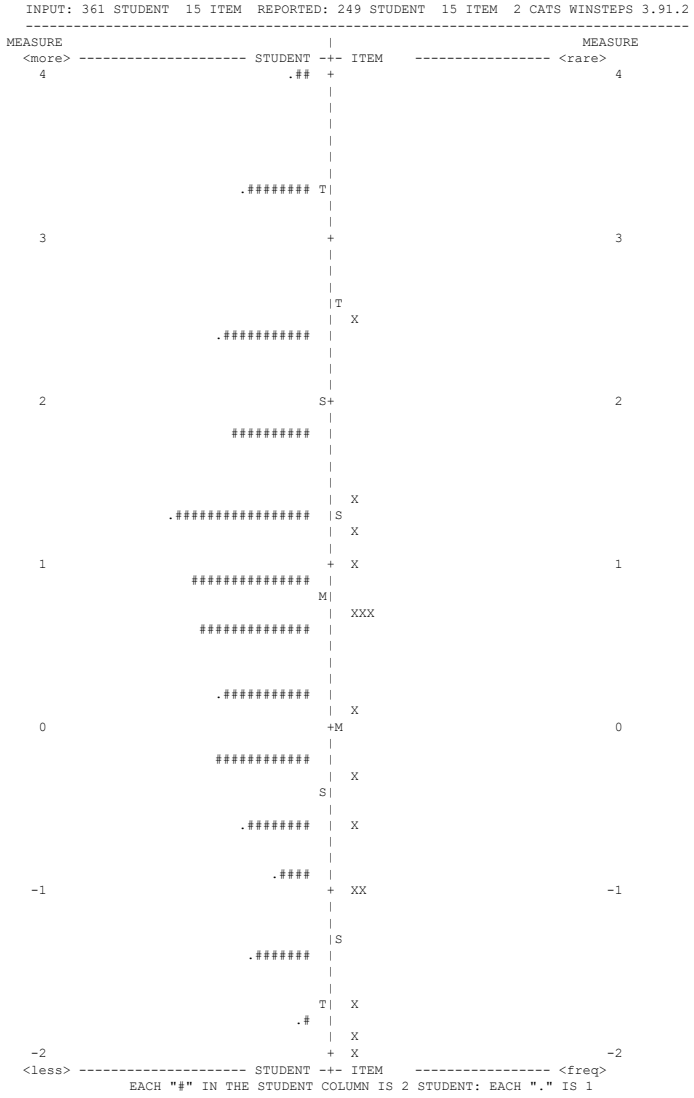

Figure 225: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2013 Lecture Test 1

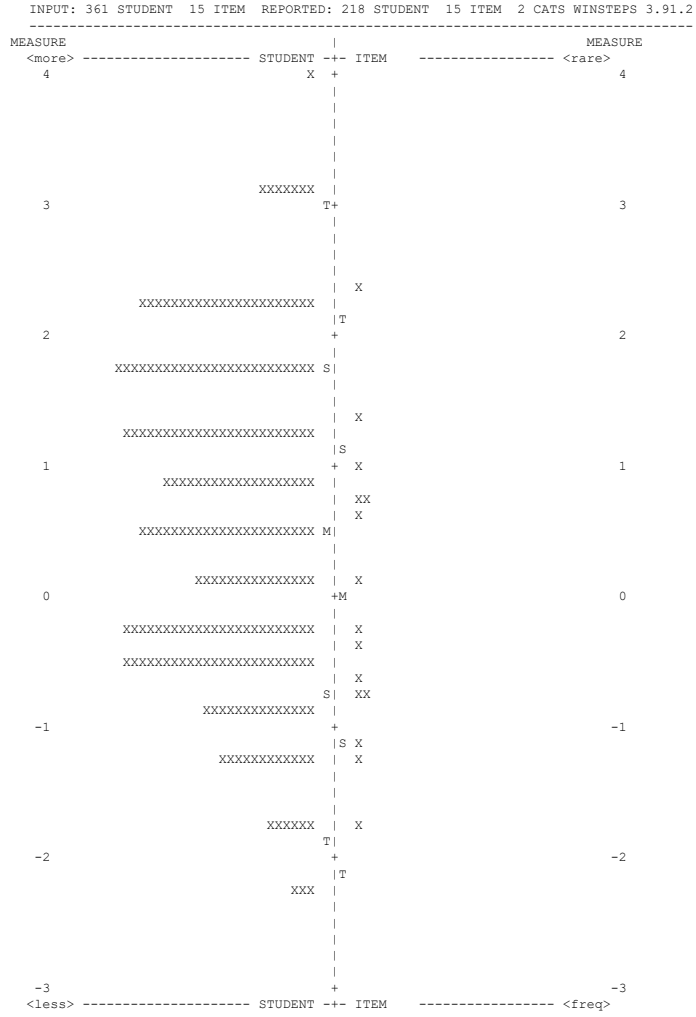

Figure 226: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2013 Lecture Test 2

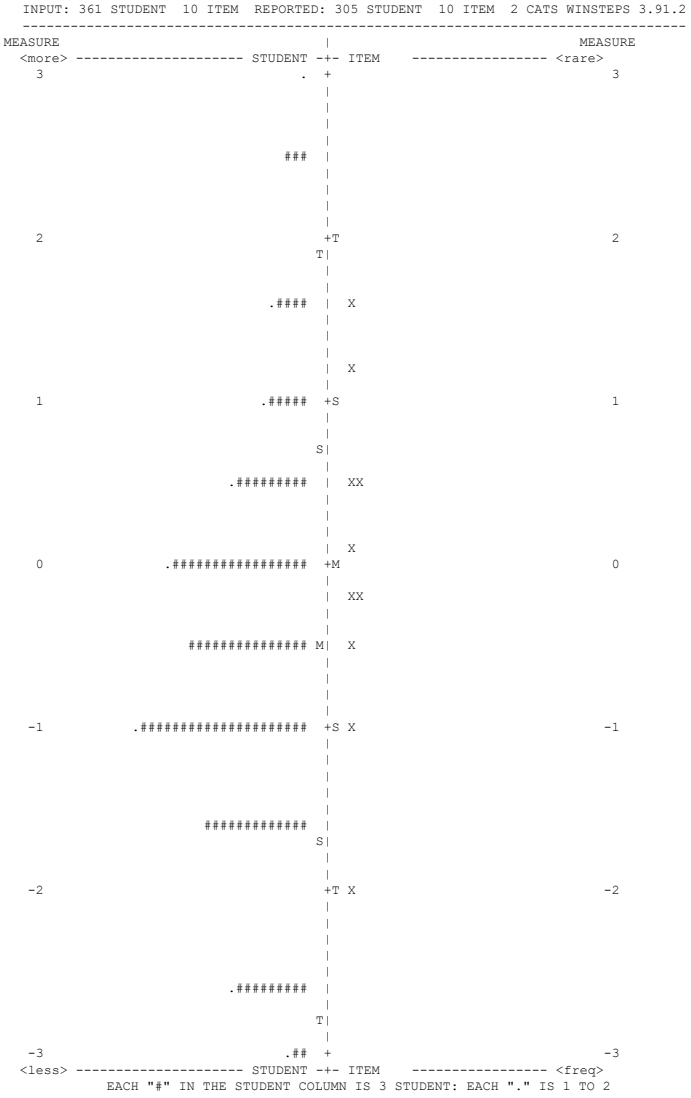

Figure 227: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2013 Exam

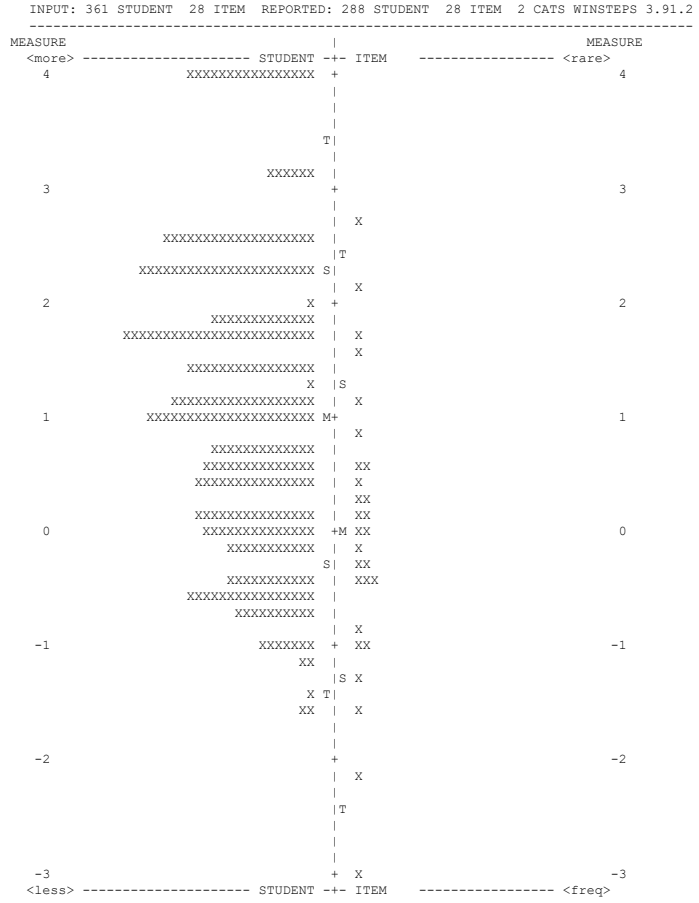

Figure 228: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2013 Redeemable Exam

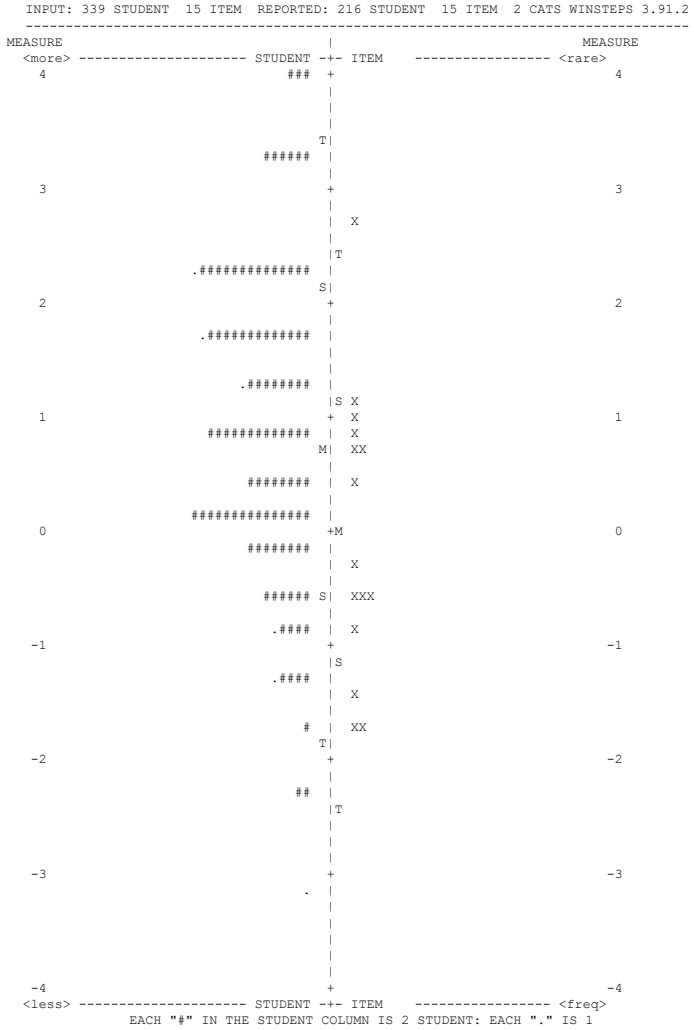

Figure 229: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2014 Lecture Test 1

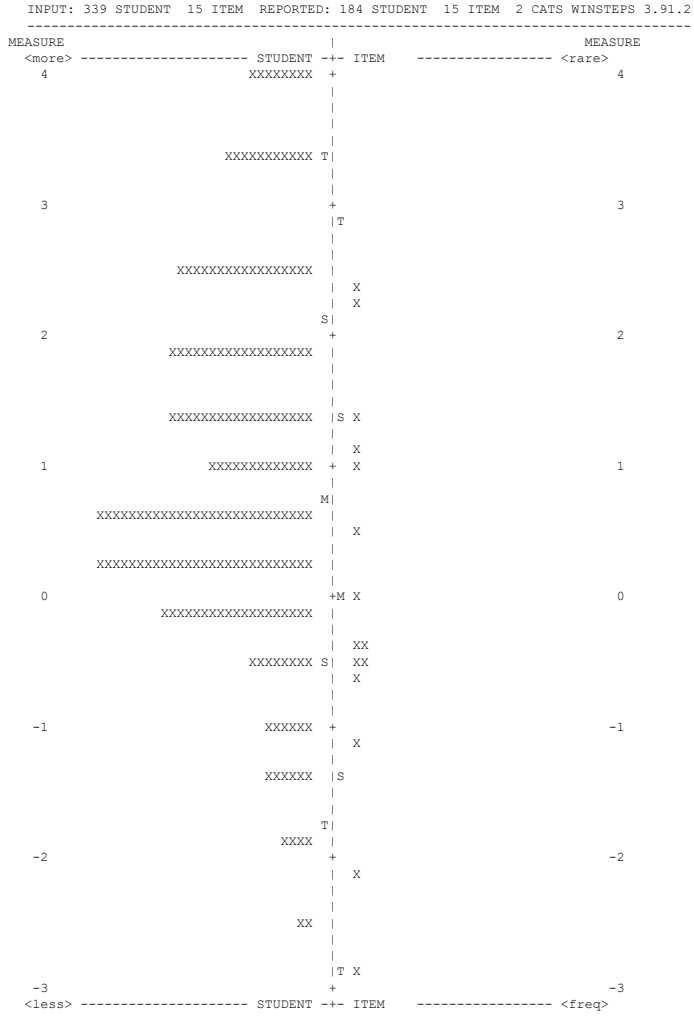

Figure 230: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2014 Lecture Test 2

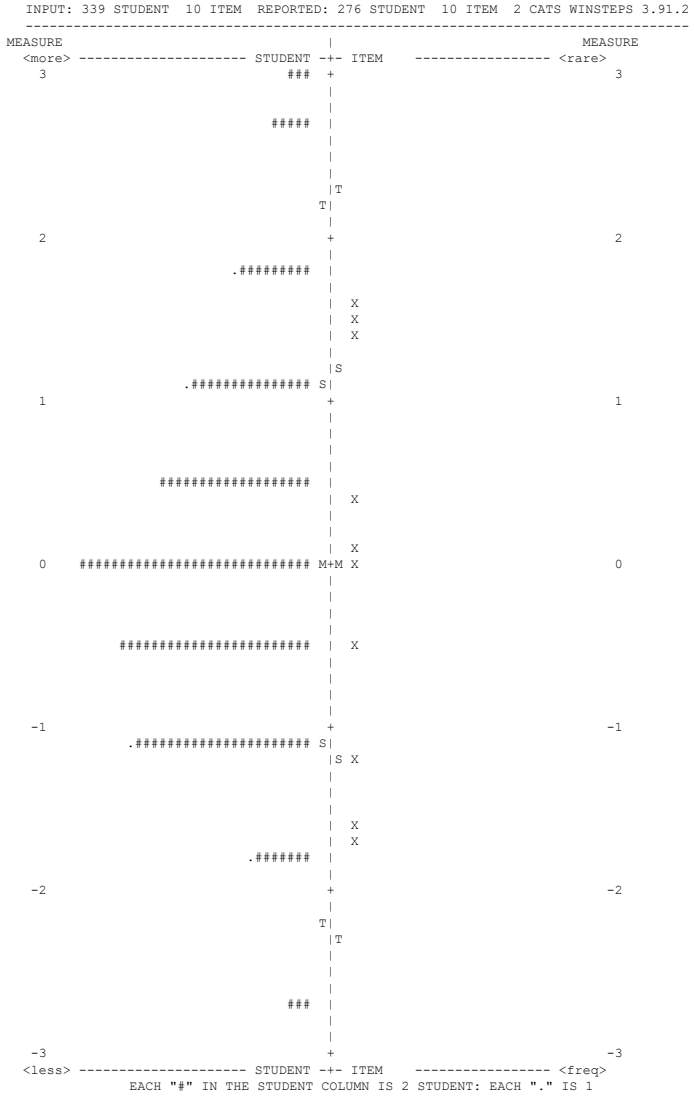

Figure 231: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2014 Exam

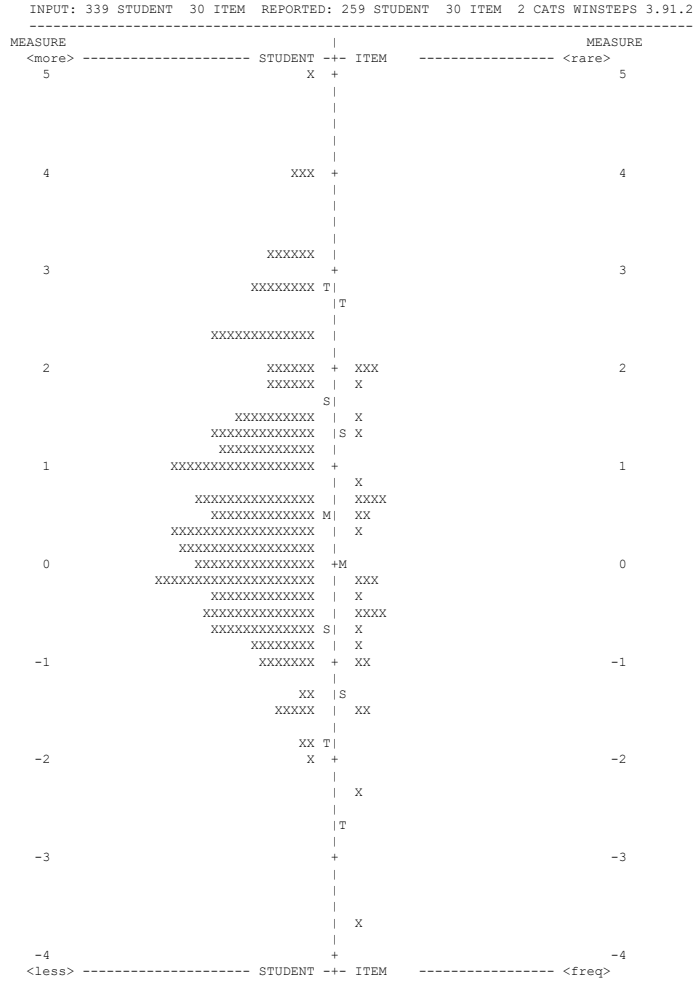

Figure 232: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2014 Redeemable Exam

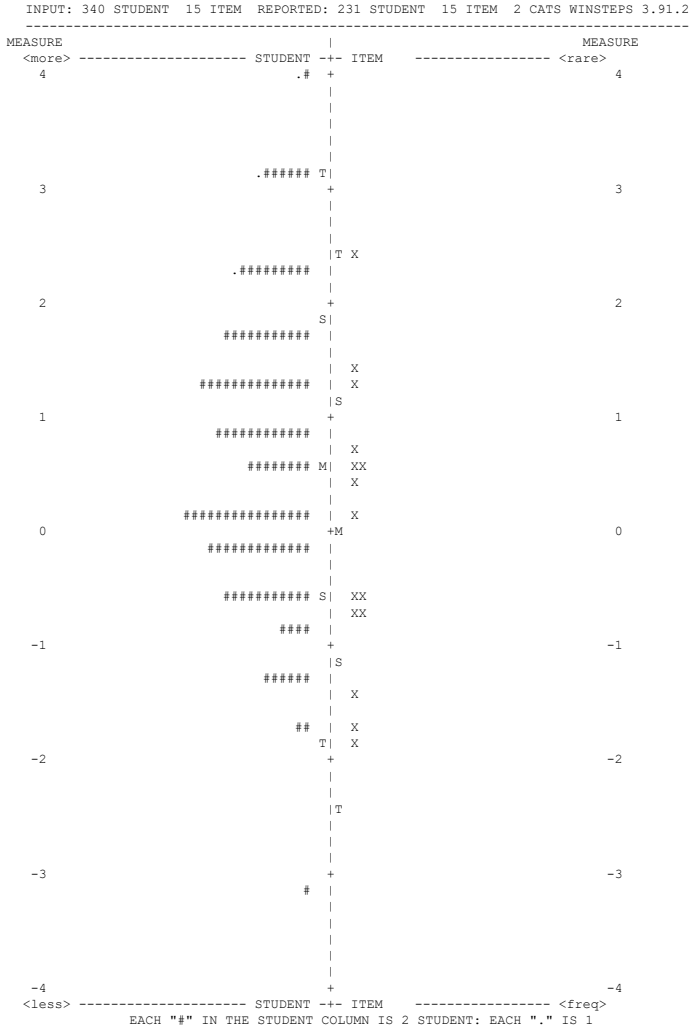

Figure 233: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2015 Lecture Test 1

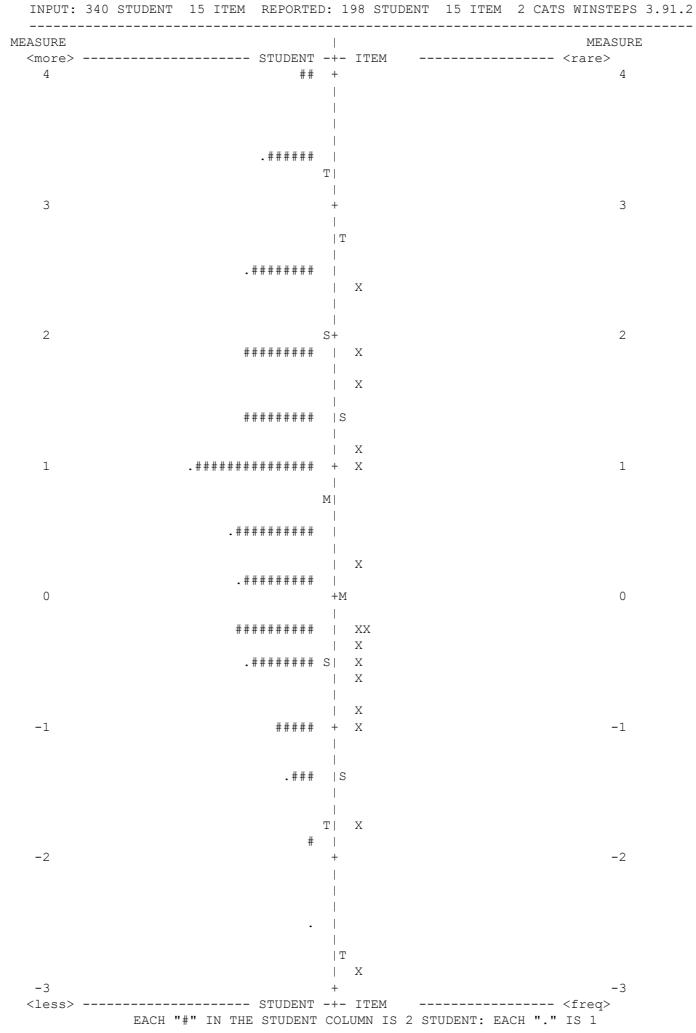

Figure 234: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2015 Lecture Test 2

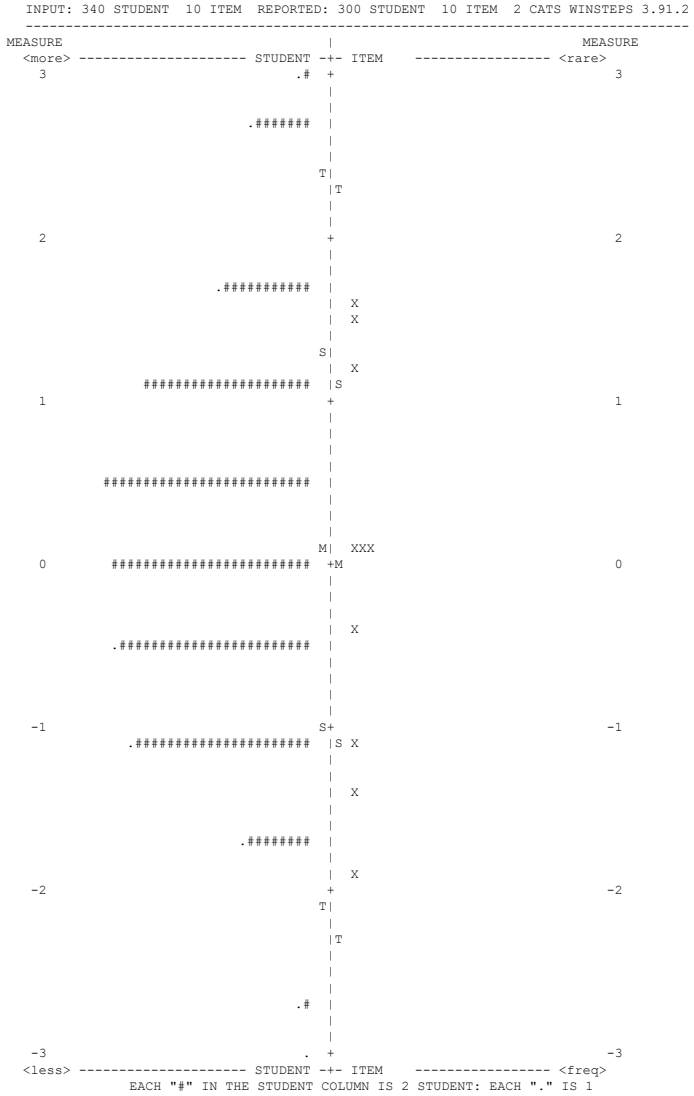

Figure 235: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2015 Exam

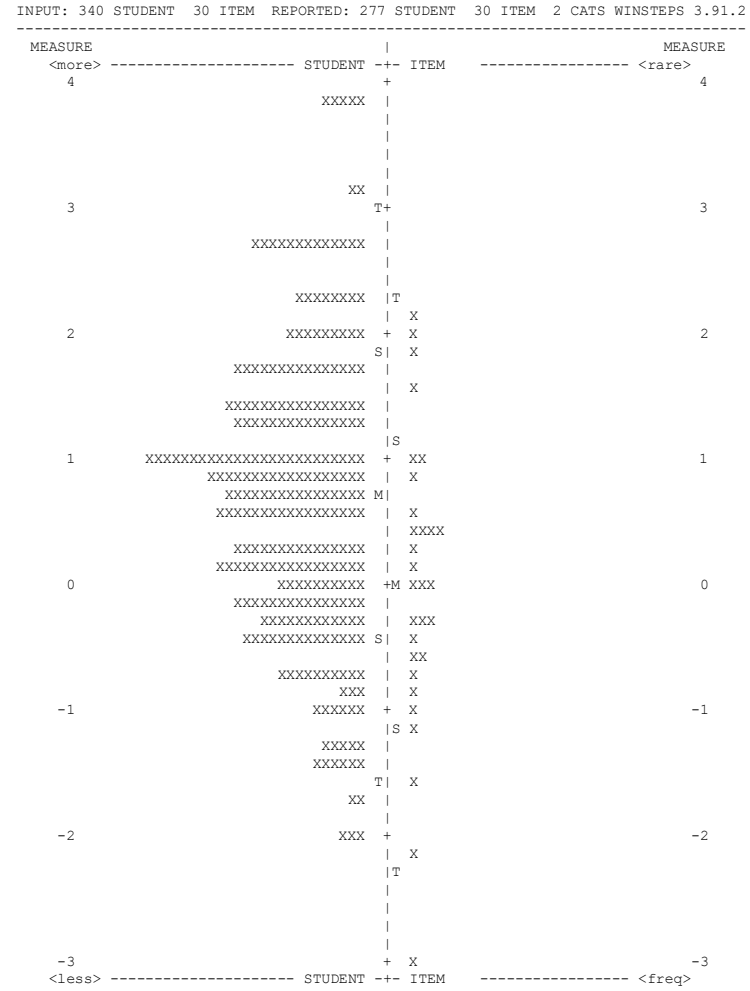

Figure 236: Wright Map of Student Ability and Item Difficulty in Foundations of Chemistry IB 2015 Redeemable Exam

7.5 MCQ Item Evaluation using CTT and Rasch Analysis

Table 49: Breakdown of Individual Items used in Chemistry IA 2012 Multiple-Choice Assessments using CTT and Rasch Analysis

Chemistry IA 2012															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	469	352	0.749	0.340	0.450	-0.91	0.12	0.96	-0.70	0.95	-0.40	0.43	0.40	77.1	77.0
Lec_1_2	469	378	0.808	0.196	0.403	-1.30	0.13	0.97	-0.40	0.99	0.00	0.39	0.37	81.1	81.4
Lec_1_3	469	198	0.421	0.553	0.508	0.89	0.11	0.96	-1.00	0.94	-0.80	0.51	0.48	74.7	70.9
Lec_1_4	469	254	0.540	0.383	0.390	0.26	0.11	1.11	2.60	1.12	1.80	0.39	0.47	62.9	69.4
Lec_1_5	469	373	0.796	0.153	0.328	-1.22	0.13	1.04	0.60	1.28	2.00	0.32	0.38	80.4	80.4
Lec_1_6	469	327	0.696	0.366	0.438	-0.58	0.11	1.00	0.00	1.01	0.10	0.42	0.42	74.5	73.7
Lec_1_7	469	240	0.511	0.553	0.522	0.42	0.11	0.94	-1.30	0.95	-0.80	0.51	0.47	72.1	69.4
Lec_1_8	469	336	0.715	0.443	0.514	-0.70	0.11	0.91	-1.70	0.78	-2.30	0.49	0.42	76.5	74.8
Lec_1_9	469	238	0.506	0.553	0.542	0.44	0.11	0.93	-1.70	0.87	-2.10	0.53	0.47	71.6	69.4
Lec_1_10	469	151	0.323	0.426	0.432	1.46	0.11	1.05	1.00	1.06	0.60	0.45	0.49	72.7	75.2
Lec_1_11	469	205	0.436	0.374	0.416	0.81	0.11	1.09	2.10	1.14	2.10	0.41	0.48	67.9	70.3
Lec_1_12	469	343	0.730	0.366	0.497	-0.79	0.12	0.92	-1.50	0.89	-1.00	0.47	0.41	75.8	75.6
Lec_1_13	469	261	0.555	0.409	0.425	0.18	0.11	1.07	1.60	1.02	0.40	0.43	0.47	64.4	69.3
Lec_1_14	469	216	0.460	0.536	0.505	0.69	0.11	0.97	-0.60	0.96	-0.60	0.50	0.48	71.0	70.0
Lec_1_15	469	245	0.521	0.477	0.443	0.36	0.11	1.04	1.10	1.08	1.20	0.44	0.47	67.9	69.4
Lec_2_1	446	262	0.587	0.619	0.423	-0.11	0.11	0.98	-0.50	0.93	-1.30	0.43	0.40	68.2	68.3
Lec_2_2	446	292	0.656	0.512	0.353	-0.45	0.11	1.04	0.90	1.04	0.70	0.35	0.39	69.1	71.0
Lec_2_3	446	255	0.572	0.646	0.459	-0.03	0.10	0.95	-1.40	0.93	-1.30	0.45	0.40	70.2	67.9
Lec_2_4	446	270	0.605	0.780	0.536	-0.20	0.11	0.87	-3.40	0.83	-2.90	0.53	0.40	74.9	68.8
Lec_2_5	446	146	0.327	0.152	0.114	1.17	0.11	1.25	5.00	1.42	5.10	0.12	0.38	66.6	71.7
Lec_2_6	446	268	0.601	0.466	0.284	-0.18	0.11	1.12	2.80	1.13	2.20	0.29	0.40	61.9	68.7
Lec_2_7	446	272	0.610	0.386	0.230	-0.22	0.11	1.17	3.90	1.20	3.20	0.24	0.40	60.1	68.9
Lec_2_8	446	100	0.224	0.395	0.307	1.78	0.12	0.98	-0.20	1.22	2.00	0.33	0.35	79.4	78.8

Lec_2_9	446	67	0.150	0.045	0.008	2.34	0.14	1.23	2.50	1.81	4.40	0.01	0.30	84.8	85.0
Lec_2_10	446	260	0.583	0.753	0.544	-0.09	0.10	0.86	-3.60	0.81	-3.60	0.54	0.40	75.3	68.2
Lec_2_11	446	356	0.798	0.529	0.353	-1.31	0.13	1.01	0.20	0.91	-0.80	0.35	0.34	79.4	80.6
Lec_2_12	446	300	0.673	0.655	0.479	-0.54	0.11	0.92	-1.70	0.84	-2.30	0.47	0.39	71.5	71.8
Lec_2_13	446	312	0.700	0.646	0.509	-0.69	0.11	0.88	-2.50	0.82	-2.40	0.50	0.38	76.9	73.5
Lec_2_14	446	302	0.677	0.628	0.484	-0.57	0.11	0.91	-1.90	0.86	-2.00	0.48	0.38	75.1	72.1
Lec_2_15	446	329	0.738	0.717	0.558	-0.91	0.12	0.83	-3.10	0.70	-3.60	0.54	0.37	79.4	76.0
Exam_1	508	334	0.657	0.370	0.333	-1.04	0.10	1.12	2.50	1.15	1.90	0.34	0.43	68.5	71.7
Exam_2	508	127	0.250	0.442	0.461	1.16	0.11	0.96	-0.60	0.91	-0.90	0.46	0.42	78.2	78.3
Exam_3	508	231	0.455	0.425	0.417	0.01	0.10	1.04	1.10	1.06	1.10	0.41	0.45	65.9	68.1
Exam_4	508	110	0.217	0.355	0.387	1.39	0.12	1.02	0.30	1.03	0.30	0.40	0.41	81.2	80.8
Exam_5	508	206	0.406	0.505	0.438	0.26	0.10	1.01	0.30	1.04	0.60	0.43	0.45	70.1	69.7
Exam_6	508	314	0.618	0.409	0.329	-0.83	0.10	1.13	3.00	1.17	2.50	0.34	0.44	65.7	70.2
Exam_7	508	343	0.675	0.512	0.489	-1.14	0.11	0.93	-1.60	0.91	-1.10	0.49	0.43	73.9	72.9
Exam_8	508	243	0.478	0.480	0.430	-0.11	0.10	1.03	0.70	1.06	1.00	0.43	0.45	67.5	68.1
Exam_9	508	209	0.412	0.536	0.554	0.23	0.10	0.88	-3.10	0.84	-2.80	0.54	0.45	76.4	69.5
Exam_10	508	225	0.444	0.568	0.565	0.07	0.10	0.87	-3.60	0.82	-3.30	0.55	0.45	71.9	68.6
Red_Exam_1	488	381	0.781	0.352	0.425	-1.06	0.12	0.94	-0.90	0.82	-1.50	0.42	0.36	81.1	79.3
Red_Exam_2	488	388	0.795	0.213	0.310	-1.17	0.12	1.03	0.50	1.16	1.10	0.31	0.35	81.3	80.5
Red_Exam_3	488	245	0.502	0.541	0.477	0.47	0.10	0.96	-1.10	0.92	-1.40	0.48	0.44	70.8	68.7
Red_Exam_4	488	335	0.686	0.549	0.551	-0.49	0.11	0.85	-3.40	0.76	-2.90	0.53	0.40	77.4	72.8
Red_Exam_5	488	385	0.789	0.172	0.274	-1.12	0.12	1.08	1.20	1.19	1.40	0.28	0.35	78.6	80.0
Red_Exam_6	488	423	0.867	0.238	0.379	-1.76	0.14	0.95	-0.60	0.75	-1.40	0.37	0.30	86.2	86.9
Red_Exam_7	488	327	0.670	0.467	0.451	-0.40	0.11	0.96	-0.90	0.94	-0.70	0.44	0.40	72.1	72.0
Red_Exam_8	488	379	0.777	0.385	0.431	-1.04	0.12	0.94	-1.00	0.80	-1.70	0.42	0.36	79.5	79.0
Red_Exam_9	488	320	0.656	0.615	0.553	-0.32	0.11	0.86	-3.50	0.80	-2.70	0.53	0.41	77.8	71.4
Red_Exam_10	488	198	0.407	0.617	0.501	0.95	0.10	0.93	-1.70	0.90	-1.60	0.50	0.44	71.7	70.6
Red_Exam_11	488	340	0.697	0.352	0.376	-0.54	0.11	1.02	0.50	1.03	0.30	0.37	0.40	74.5	73.4
Red_Exam_12	488	311	0.639	0.468	0.425	-0.22	0.10	1.00	-0.10	0.98	-0.30	0.42	0.41	71.9	70.7
Red_Exam_13	488	240	0.492	0.500	0.407	0.52	0.10	1.03	0.90	1.00	0.10	0.42	0.44	65.3	68.8

Red_Exam_14	488	309	0.634	0.542	0.466	-0.20	0.10	0.95	-1.10	0.92	-1.00	0.45	0.42	74.7	70.5
Red_Exam_15	488	203	0.416	0.516	0.440	0.90	0.10	0.98	-0.40	1.00	0.00	0.45	0.44	71.5	70.3
Red_Exam_16	488	272	0.557	0.500	0.465	0.19	0.10	0.97	-0.80	0.97	-0.50	0.46	0.43	69.8	68.9
Red_Exam_17	488	320	0.656	0.402	0.373	-0.32	0.11	1.05	1.10	1.04	0.50	0.37	0.41	70.0	71.4
Red_Exam_18	488	227	0.465	0.631	0.554	0.65	0.10	0.88	-3.20	0.84	-2.80	0.54	0.44	73.7	69.1
Red_Exam_19	488	254	0.520	0.320	0.262	0.37	0.10	1.19	4.80	1.25	4.00	0.28	0.44	58.7	68.7
Red_Exam_20	488	318	0.652	0.426	0.430	-0.30	0.10	0.99	-0.30	0.97	-0.30	0.42	0.41	72.1	71.2
Red_Exam_21	488	291	0.596	0.631	0.572	-0.01	0.10	0.85	-4.00	0.78	-3.40	0.55	0.42	76.2	69.6
Red_Exam_22	488	267	0.547	0.172	0.179	0.24	0.10	1.28	6.80	1.54	7.60	0.18	0.43	58.1	68.8
Red_Exam_23	488	122	0.250	0.279	0.285	1.85	0.12	1.11	1.80	1.33	3.00	0.31	0.42	77.4	78.4
Red_Exam_24	488	56	0.115	-0.041	-0.038	2.98	0.15	1.34	3.00	2.70	6.20	-0.02	0.34	86.4	88.9
Red_Exam_25	488	317	0.650	0.525	0.522	-0.29	0.10	0.90	-2.50	0.80	-2.60	0.51	0.41	72.7	71.1
Red_Exam_26	488	367	0.754	0.255	0.308	-0.88	0.11	1.07	1.20	1.08	0.80	0.31	0.37	75.8	77.1
Red_Exam_27	488	166	0.340	0.557	0.447	1.30	0.11	0.95	-1.10	1.02	0.40	0.47	0.44	77.0	73.3
Red_Exam_28	488	332	0.680	0.377	0.397	-0.45	0.11	1.02	0.40	0.95	-0.60	0.40	0.40	70.6	72.5
Red_Exam_29	488	338	0.693	0.467	0.443	-0.52	0.11	0.96	-0.80	0.92	-0.80	0.43	0.40	73.3	73.1
Red_Exam_30	488	228	0.467	0.541	0.452	0.64	0.10	0.98	-0.40	0.96	-0.60	0.46	0.44	69.0	69.1

Table 50: Breakdown of Individual Items used in Chemistry IA 2013 Multiple-Choice Assessments using CTT and Rasch Analysis

Chemistry IA 2013															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	448	367	0.821	0.500	0.441	-1.29	0.13	0.95	-0.60	0.84	-1.00	0.41	0.37	82.6	82.3
Lec_1_2	448	344	0.774	0.493	0.374	-0.91	0.12	1.04	0.70	1.05	0.40	0.37	0.40	77.5	78.2
Lec_1_3	448	210	0.468	0.633	0.431	0.78	0.11	1.06	1.30	1.17	2.50	0.44	0.49	68.1	70.3
Lec_1_4	448	225	0.501	0.686	0.481	0.61	0.11	1.01	0.20	0.99	-0.20	0.49	0.49	69.7	70.1
Lec_1_5	448	343	0.766	0.490	0.384	-0.89	0.12	1.04	0.60	1.06	0.50	0.37	0.40	77.8	78.0
Lec_1_6	448	286	0.639	0.650	0.447	-0.12	0.11	1.02	0.60	1.12	1.50	0.43	0.46	71.3	71.4
Lec_1_7	448	231	0.517	0.775	0.554	0.54	0.11	0.92	-1.90	0.90	-1.60	0.54	0.49	74.3	70.0
Lec_1_8	448	337	0.753	0.650	0.490	-0.80	0.12	0.92	-1.30	0.82	-1.70	0.47	0.41	78.7	77.0
Lec_1_9	448	223	0.497	0.766	0.530	0.63	0.11	0.94	-1.40	0.95	-0.80	0.53	0.49	74.3	70.2
Lec_1_10	448	152	0.342	0.580	0.438	1.51	0.12	1.06	1.10	1.07	0.80	0.46	0.50	71.3	74.3
Lec_1_11	448	218	0.488	0.668	0.483	0.69	0.11	1.01	0.40	1.00	0.00	0.48	0.49	69.9	70.2
Lec_1_12	448	358	0.800	0.543	0.446	-1.13	0.13	0.94	-0.90	0.84	-1.10	0.43	0.38	82.6	80.5
Lec_1_13	448	301	0.673	0.722	0.505	-0.31	0.11	0.94	-1.20	0.93	-0.80	0.49	0.45	75.5	72.6
Lec_1_14	448	243	0.542	0.661	0.494	0.39	0.11	0.99	-0.30	1.03	0.40	0.48	0.48	72.2	70.0
Lec_1_15	448	249	0.557	0.570	0.402	0.32	0.11	1.11	2.40	1.14	2.10	0.40	0.48	64.8	70.1
Lec_2_1	420	289	0.689	0.627	0.442	-0.48	0.12	0.97	-0.60	0.91	-1.00	0.44	0.40	72.6	73.3
Lec_2_2	420	284	0.677	0.675	0.443	-0.41	0.11	0.97	-0.50	0.92	-1.00	0.44	0.41	72.9	72.7
Lec_2_3	420	259	0.618	0.760	0.502	-0.10	0.11	0.92	-1.90	0.90	-1.60	0.49	0.42	74.3	70.3
Lec_2_4	420	282	0.672	0.827	0.532	-0.39	0.11	0.88	-2.50	0.83	-2.20	0.52	0.41	74.8	72.5
Lec_2_5	420	112	0.266	-0.019	0.010	1.70	0.12	1.33	5.30	1.84	6.70	0.03	0.38	73.3	75.0
Lec_2_6	420	241	0.572	0.513	0.298	0.12	0.11	1.13	3.00	1.15	2.30	0.31	0.42	61.9	69.2
Lec_2_7	420	280	0.667	0.418	0.235	-0.36	0.11	1.19	3.70	1.23	2.70	0.24	0.41	65.2	72.2
Lec_2_8	420	127	0.302	0.589	0.391	1.50	0.12	0.98	-0.50	0.98	-0.20	0.41	0.39	75.5	73.3
Lec_2_9	420	50	0.119	0.029	0.027	2.84	0.16	1.22	1.90	2.03	4.10	0.02	0.29	86.9	88.3
Lec_2_10	420	259	0.618	0.846	0.558	-0.10	0.11	0.86	-3.30	0.80	-3.10	0.55	0.42	75.7	70.3
Lec_2_11	420	337	0.803	0.428	0.327	-1.20	0.13	1.05	0.80	0.99	-0.10	0.32	0.36	81.2	81.3

Lec_2_12	420	283	0.675	0.694	0.455	-0.40	0.11	0.96	-0.70	0.88	-1.50	0.45	0.41	71.2	72.6
Lec_2_13	420	306	0.727	0.741	0.535	-0.71	0.12	0.86	-2.40	0.74	-2.80	0.52	0.39	78.8	75.8
Lec_2_14	420	337	0.805	0.600	0.466	-1.20	0.13	0.91	-1.20	0.79	-1.70	0.45	0.36	81.2	81.3
Lec_2_15	420	313	0.743	0.827	0.602	-0.82	0.12	0.79	-3.60	0.64	-4.00	0.58	0.39	81.0	76.9
Exam_1	505	367	0.727	0.609	0.300	-1.35	0.11	1.12	2.20	1.24	2.50	0.30	0.41	71.9	76.1
Exam_2	505	114	0.225	0.522	0.395	1.38	0.12	1.01	0.10	0.99	-0.10	0.40	0.40	80.2	79.8
Exam_3	505	247	0.489	0.657	0.373	-0.09	0.10	1.07	2.00	1.08	1.40	0.38	0.44	64.0	67.6
Exam_4	505	137	0.273	0.458	0.312	1.08	0.11	1.12	2.10	1.18	1.80	0.32	0.41	73.9	76.1
Exam_5	505	214	0.423	0.648	0.393	0.24	0.10	1.05	1.30	1.04	0.70	0.40	0.44	66.3	68.7
Exam_6	505	274	0.543	0.680	0.388	-0.35	0.10	1.06	1.60	1.08	1.40	0.39	0.44	67.1	68.3
Exam_7	505	301	0.600	1.026	0.546	-0.62	0.10	0.88	-3.10	0.82	-3.10	0.53	0.43	74.1	69.6
Exam_8	505	292	0.579	0.949	0.528	-0.53	0.10	0.90	-2.60	0.85	-2.70	0.52	0.43	71.1	68.9
Exam_9	505	213	0.423	0.854	0.488	0.25	0.10	0.95	-1.40	0.93	-1.30	0.48	0.44	70.1	68.7
Exam_10	505	240	0.477	0.990	0.555	-0.02	0.10	0.87	-3.60	0.82	-3.50	0.54	0.44	71.9	67.4
Red_Exam_1	504	376	0.747	0.577	0.510	-0.99	0.12	0.95	-0.80	0.85	-1.30	0.50	0.46	79.3	78.5
Red_Exam_2	504	385	0.768	0.349	0.381	-1.12	0.12	1.09	1.40	1.22	1.60	0.40	0.46	79.1	80.0
Red_Exam_3	504	261	0.518	0.735	0.515	0.30	0.10	0.93	-1.80	0.90	-1.70	0.50	0.45	72.1	68.6
Red_Exam_4	504	336	0.666	0.664	0.547	-0.50	0.11	0.90	-2.30	1.00	0.00	0.52	0.46	76.0	73.2
Red_Exam_5	504	396	0.788	0.356	0.372	-1.28	0.12	1.07	1.00	1.32	2.10	0.40	0.46	82.6	81.9
Red_Exam_6	504	419	0.834	0.539	0.553	-1.68	0.14	0.86	-1.60	0.73	-1.60	0.55	0.47	87.1	86.1
Red_Exam_7	504	352	0.701	0.657	0.537	-0.69	0.11	0.92	-1.50	0.83	-1.70	0.52	0.46	76.2	75.1
Red_Exam_8	504	351	0.699	0.467	0.428	-0.67	0.11	1.05	0.90	1.10	1.00	0.43	0.46	75.6	75.0
Red_Exam_9	504	335	0.664	0.711	0.550	-0.49	0.11	0.91	-2.00	0.90	-1.20	0.52	0.46	77.0	73.1
Red_Exam_10	504	194	0.383	0.672	0.485	1.00	0.10	0.95	-1.30	0.93	-1.10	0.47	0.43	73.2	71.0
Red_Exam_11	504	354	0.704	0.482	0.443	-0.71	0.11	1.04	0.80	0.98	-0.10	0.44	0.46	73.4	75.4
Red_Exam_12	504	280	0.555	0.632	0.453	0.11	0.10	1.02	0.60	1.00	0.00	0.44	0.46	66.6	69.0
Red_Exam_13	504	258	0.513	0.610	0.451	0.33	0.10	1.02	0.50	0.99	-0.10	0.44	0.45	67.0	68.6
Red_Exam_14	504	339	0.674	0.672	0.531	-0.53	0.11	0.92	-1.70	0.87	-1.40	0.51	0.46	75.0	73.5
Red_Exam_15	504	232	0.460	0.751	0.533	0.60	0.10	0.91	-2.40	0.90	-1.70	0.50	0.45	72.5	69.1
Red_Exam_16	504	282	0.559	0.664	0.487	0.09	0.10	0.98	-0.40	0.96	-0.60	0.47	0.46	69.5	69.1

Red_Exam_17	504	300	0.595	0.648	0.521	-0.10	0.10	0.93	-1.60	0.90	-1.40	0.50	0.46	72.3	70.0
Red_Exam_18	504	210	0.416	0.634	0.479	0.83	0.10	0.97	-0.80	0.98	-0.30	0.46	0.44	71.1	69.9
Red_Exam_19	504	253	0.502	0.411	0.312	0.39	0.10	1.19	4.80	1.24	3.60	0.33	0.45	59.8	68.7
Red_Exam_20	504	308	0.611	0.648	0.520	-0.19	0.10	0.94	-1.50	0.94	-0.80	0.49	0.46	72.7	70.5
Red_Exam_21	504	290	0.577	0.791	0.567	0.00	0.10	0.89	-3.00	0.85	-2.30	0.53	0.46	73.6	69.5
Red_Exam_22	504	212	0.421	0.166	0.163	0.81	0.10	1.37	8.20	1.58	8.10	0.19	0.44	55.9	69.8
Red_Exam_23	504	174	0.348	0.561	0.448	1.22	0.11	0.97	-0.60	1.00	0.10	0.44	0.42	74.8	72.7
Red_Exam_24	504	66	0.130	0.672	0.505	2.76	0.14	1.32	3.20	2.41	5.90	0.06	0.33	86.3	87.2
Red_Exam_25	504	321	0.638	0.688	0.546	-0.33	0.11	0.92	-2.00	0.83	-2.10	0.52	0.46	73.8	71.6
Red_Exam_26	504	346	0.687	0.602	0.496	-0.61	0.11	0.97	-0.50	0.91	-0.90	0.48	0.46	76.4	74.4
Red_Exam_27	504	152	0.300	0.498	0.385	1.47	0.11	1.00	0.00	1.12	1.50	0.40	0.41	76.0	74.9
Red_Exam_28	504	331	0.658	0.632	0.501	-0.44	0.11	0.97	-0.70	0.99	0.00	0.48	0.46	74.6	72.6
Red_Exam_29	504	367	0.729	0.609	0.530	-0.88	0.11	0.92	-1.50	0.91	-0.80	0.51	0.46	78.7	77.2
Red_Exam_30	504	166	0.330	0.506	0.405	1.31	0.11	1.00	0.00	1.08	1.10	0.41	0.42	75.2	73.5

Table 51: Breakdown of Individual Items used in Chemistry IA 2014 Multiple-Choice Assessments using CTT and Rasch Analysis

Chemistry IA 2014															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	474	148	0.314	0.510	0.449	1.29	0.11	0.98	-0.40	1.03	0.30	0.46	0.45	77.4	74.8
Lec_1_2	474	299	0.632	0.490	0.452	-0.37	0.11	0.97	-0.60	0.99	-0.20	0.44	0.42	73.8	70.8
Lec_1_3	474	388	0.819	0.228	0.369	-1.54	0.13	0.99	-0.10	1.06	0.40	0.35	0.34	81.7	82.6
Lec_1_4	474	387	0.816	0.245	0.404	-1.52	0.13	0.93	-0.90	0.96	-0.20	0.39	0.35	83.6	82.4
Lec_1_5	474	284	0.599	0.439	0.457	-0.21	0.10	0.98	-0.40	0.94	-0.80	0.45	0.43	70.9	69.6
Lec_1_6	474	281	0.593	0.532	0.492	-0.18	0.10	0.94	-1.60	1.04	0.60	0.47	0.43	74.0	69.5
Lec_1_7	474	89	0.188	0.338	0.378	2.14	0.13	0.99	-0.10	1.10	0.80	0.41	0.42	84.3	83.6
Lec_1_8	474	249	0.525	0.489	0.449	0.17	0.10	1.00	0.10	0.97	-0.50	0.45	0.44	68.3	68.6
Lec_1_9	474	115	0.243	0.321	0.331	1.73	0.12	1.12	1.80	1.15	1.30	0.35	0.43	75.5	79.2
Lec_1_10	474	327	0.690	0.464	0.474	-0.70	0.11	0.94	-1.30	0.86	-1.50	0.46	0.40	74.9	73.4
Lec_1_11	474	376	0.795	0.245	0.354	-1.35	0.12	1.02	0.40	1.14	1.00	0.34	0.36	79.1	80.3
Lec_1_12	474	149	0.314	0.549	0.481	1.28	0.11	0.93	-1.40	0.96	-0.50	0.49	0.45	77.7	74.7
Lec_1_13	474	209	0.441	0.532	0.436	0.59	0.10	1.01	0.20	1.01	0.10	0.44	0.45	70.0	69.7
Lec_1_14	474	300	0.634	0.288	0.327	-0.39	0.11	1.11	2.60	1.28	3.30	0.32	0.42	66.8	70.9
Lec_1_15	474	347	0.732	0.304	0.400	-0.95	0.11	1.00	0.00	1.01	0.10	0.39	0.39	75.5	75.8
Lec_2_1	436	269	0.617	0.523	0.462	-0.03	0.11	0.96	-1.00	0.93	-1.00	0.46	0.42	69.7	70.3
Lec_2_2	436	301	0.690	0.413	0.343	-0.42	0.11	1.07	1.40	1.07	0.90	0.34	0.41	70.3	73.7
Lec_2_3	436	220	0.505	0.688	0.578	0.53	0.11	0.83	-4.60	0.79	-3.90	0.56	0.42	75.9	67.6
Lec_2_4	436	289	0.664	0.616	0.546	-0.27	0.11	0.87	-2.80	0.81	-2.70	0.53	0.41	76.8	72.2
Lec_2_5	436	118	0.271	0.055	0.080	1.74	0.12	1.27	4.50	1.61	5.00	0.09	0.37	71.0	75.4
Lec_2_6	436	254	0.583	0.394	0.312	0.15	0.11	1.11	2.70	1.12	1.90	0.32	0.42	64.1	69.1
Lec_2_7	436	284	0.651	0.349	0.291	-0.21	0.11	1.13	2.60	1.23	3.00	0.29	0.41	68.7	71.6
Lec_2_8	436	136	0.312	0.394	0.351	1.50	0.11	1.02	0.40	1.01	0.10	0.36	0.38	70.6	72.5
Lec_2_9	436	234	0.537	0.550	0.402	0.37	0.11	1.01	0.40	1.07	1.30	0.40	0.42	65.1	67.9
Lec_2_10	436	244	0.560	0.550	0.432	0.26	0.11	0.99	-0.20	0.96	-0.70	0.43	0.42	68.7	68.5
Lec_2_11	436	350	0.803	0.339	0.330	-1.14	0.13	1.05	0.70	1.04	0.40	0.33	0.37	81.8	81.7

Lec_2_12	436	295	0.677	0.615	0.509	-0.34	0.11	0.90	-1.90	0.89	-1.50	0.49	0.41	76.8	73.0
Lec_2_13	436	307	0.706	0.469	0.369	-0.50	0.12	1.04	0.80	1.04	0.50	0.36	0.40	73.8	74.4
Lec_2_14	436	322	0.739	0.532	0.497	-0.70	0.12	0.91	-1.60	0.81	-2.10	0.49	0.39	78.6	76.8
Lec_2_15	436	339	0.778	0.615	0.585	-0.96	0.13	0.80	-3.10	0.66	-3.30	0.56	0.38	82.5	79.7
Exam_1	508	235	0.462	0.574	0.422	0.76	0.10	1.11	2.50	1.22	3.30	0.43	0.51	67.4	70.5
Exam_2	508	372	0.731	0.629	0.533	-0.74	0.11	0.88	-2.30	0.78	-2.40	0.51	0.43	80.0	75.8
Exam_3	508	384	0.756	0.495	0.418	-0.90	0.11	1.02	0.30	1.00	0.00	0.41	0.41	75.8	77.1
Exam_4	508	330	0.650	0.770	0.605	-0.25	0.11	0.82	-4.20	0.76	-3.50	0.58	0.46	77.9	71.8
Exam_5	508	327	0.642	0.338	0.268	-0.22	0.11	1.26	5.60	1.38	4.70	0.28	0.46	62.0	71.6
Exam_6	508	374	0.737	0.448	0.393	-0.77	0.11	1.04	0.90	1.10	1.00	0.39	0.42	75.4	76.0
Exam_7	508	362	0.711	0.629	0.525	-0.62	0.11	0.90	-2.10	0.91	-1.00	0.50	0.44	78.3	74.6
Exam_8	508	265	0.523	0.747	0.512	0.44	0.10	0.98	-0.50	0.99	-0.10	0.51	0.50	70.0	69.3
Exam_9	508	177	0.348	0.613	0.467	1.41	0.11	1.03	0.50	1.10	1.10	0.49	0.51	74.0	74.9
Exam_10	508	223	0.438	0.762	0.561	0.89	0.10	0.91	-2.00	0.89	-1.70	0.57	0.51	74.4	70.9
Red_Exam_1	509	210	0.413	0.582	0.429	0.86	0.10	1.00	-0.10	0.99	-0.10	0.42	0.42	68.7	69.1
Red_Exam_2	509	397	0.780	0.511	0.416	-1.15	0.12	1.01	0.20	0.95	-0.30	0.41	0.42	79.7	80.4
Red_Exam_3	509	318	0.625	0.621	0.461	-0.21	0.10	0.98	-0.60	0.98	-0.30	0.44	0.43	70.5	70.1
Red_Exam_4	509	255	0.501	0.660	0.453	0.41	0.10	0.99	-0.40	1.02	0.30	0.43	0.43	68.5	67.8
Red_Exam_5	509	372	0.731	0.527	0.448	-0.82	0.11	0.97	-0.60	0.97	-0.20	0.44	0.42	77.5	76.3
Red_Exam_6	509	308	0.605	0.699	0.528	-0.11	0.10	0.90	-2.50	0.83	-2.60	0.50	0.43	73.1	69.4
Red_Exam_7	509	160	0.314	0.448	0.360	1.39	0.11	1.04	0.80	1.13	1.70	0.36	0.40	73.3	73.6
Red_Exam_8	509	282	0.554	0.558	0.426	0.15	0.10	1.02	0.60	1.02	0.30	0.41	0.43	68.1	68.1
Red_Exam_9	509	144	0.283	0.369	0.318	1.57	0.11	1.10	1.80	1.12	1.40	0.32	0.39	72.5	75.5
Red_Exam_10	509	380	0.747	0.629	0.507	-0.92	0.11	0.92	-1.40	0.79	-2.00	0.49	0.42	79.1	77.6
Red_Exam_11	509	376	0.739	0.377	0.329	-0.87	0.11	1.08	1.40	1.43	3.60	0.34	0.42	77.1	77.0
Red_Exam_12	509	268	0.527	0.644	0.463	0.29	0.10	0.97	-0.80	1.01	0.20	0.44	0.43	70.5	67.8
Red_Exam_13	509	319	0.627	0.762	0.556	-0.22	0.10	0.87	-3.30	0.79	-3.10	0.53	0.43	75.1	70.2
Red_Exam_14	509	411	0.807	0.440	0.366	-1.36	0.13	1.03	0.40	1.20	1.40	0.38	0.41	82.9	82.9
Red_Exam_15	509	425	0.835	0.511	0.473	-1.59	0.13	0.93	-0.80	0.76	-1.60	0.47	0.41	85.9	85.5
Red_Exam_16	509	288	0.566	0.621	0.431	0.09	0.10	1.01	0.40	1.01	0.20	0.42	0.43	67.7	68.4

Red_Exam_17	509	315	0.619	0.542	0.384	-0.18	0.10	1.06	1.60	1.08	1.10	0.38	0.43	67.1	69.9
Red_Exam_18	509	186	0.365	0.621	0.428	1.11	0.10	1.01	0.10	1.00	0.00	0.41	0.41	72.1	71.0
Red_Exam_19	509	271	0.532	0.487	0.346	0.26	0.10	1.11	2.90	1.12	2.10	0.35	0.43	63.9	67.9
Red_Exam_20	509	335	0.658	0.621	0.457	-0.39	0.10	0.98	-0.40	0.91	-1.10	0.45	0.43	71.9	71.6
Red_Exam_21	509	318	0.625	0.794	0.591	-0.21	0.10	0.83	-4.40	0.74	-3.90	0.55	0.43	76.5	70.1
Red_Exam_22	509	242	0.475	0.236	0.158	0.54	0.10	1.33	8.20	1.41	6.70	0.18	0.42	52.2	67.9
Red_Exam_23	509	239	0.470	0.638	0.396	0.57	0.10	1.04	1.20	1.06	1.20	0.39	0.42	65.7	67.9
Red_Exam_24	509	272	0.534	0.550	0.408	0.25	0.10	1.04	1.20	1.11	1.80	0.39	0.43	66.9	67.9
Red_Exam_25	509	340	0.668	0.692	0.506	-0.45	0.10	0.92	-1.80	0.91	-1.10	0.48	0.43	75.7	72.1
Red_Exam_26	509	368	0.723	0.613	0.468	-0.77	0.11	0.96	-0.80	0.90	-1.00	0.45	0.42	77.1	75.7
Red_Exam_27	509	148	0.291	0.574	0.470	1.52	0.11	0.90	-1.90	0.88	-1.50	0.46	0.40	77.7	75.0
Red_Exam_28	509	342	0.672	0.629	0.437	-0.47	0.11	1.00	0.10	0.95	-0.60	0.43	0.42	72.5	72.3
Red_Exam_29	509	366	0.720	0.701	0.521	-0.75	0.11	0.90	-2.00	0.81	-2.00	0.50	0.42	78.3	75.5
Red_Exam_30	509	153	0.301	0.479	0.379	1.47	0.11	1.00	0.10	1.11	1.40	0.38	0.40	76.7	74.4

Table 52: Breakdown of Individual Items used in Chemistry IA 2015 Multiple-Choice Assessments using CTT and Rasch Analysis

Chemistry IA 2015															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	504	153	0.304	0.452	0.347	1.49	0.11	1.11	2.00	1.27	2.80	0.37	0.46	72.3	75.8
Lec_1_2	504	337	0.669	0.619	0.464	-0.44	0.10	0.95	-1.10	0.93	-0.90	0.44	0.41	74.5	72.1
Lec_1_3	504	417	0.827	0.468	0.388	-1.47	0.13	0.96	-0.50	0.85	-1.10	0.37	0.34	84.6	83.2
Lec_1_4	504	415	0.823	0.571	0.468	-1.44	0.13	0.87	-1.90	0.92	-0.50	0.44	0.34	84.6	82.8
Lec_1_5	504	278	0.552	0.627	0.467	0.18	0.10	0.97	-0.70	0.96	-0.70	0.46	0.44	70.4	68.2
Lec_1_6	504	320	0.635	0.643	0.425	-0.25	0.10	1.01	0.30	0.98	-0.20	0.42	0.42	70.6	70.5
Lec_1_7	504	123	0.244	0.587	0.501	1.88	0.12	0.88	-2.00	0.89	-1.00	0.53	0.45	82.2	79.6
Lec_1_8	504	293	0.581	0.532	0.381	0.03	0.10	1.08	2.00	1.07	1.10	0.38	0.43	65.4	68.9
Lec_1_9	504	121	0.240	0.429	0.379	1.91	0.12	1.04	0.70	1.20	1.70	0.41	0.45	81.0	79.9
Lec_1_10	504	371	0.736	0.619	0.445	-0.83	0.11	0.96	-0.70	0.87	-1.40	0.43	0.39	75.9	76.1
Lec_1_11	504	402	0.798	0.468	0.366	-1.24	0.12	0.99	-0.10	1.11	0.90	0.35	0.35	81.2	80.6
Lec_1_12	504	188	0.373	0.651	0.459	1.10	0.10	1.00	0.00	1.03	0.40	0.46	0.46	73.3	72.0
Lec_1_13	504	236	0.470	0.629	0.464	0.60	0.10	0.99	-0.20	0.96	-0.80	0.46	0.45	67.8	68.5
Lec_1_14	504	348	0.690	0.476	0.334	-0.56	0.11	1.08	1.80	1.31	3.40	0.32	0.40	70.9	73.1
Lec_1_15	504	382	0.758	0.516	0.391	-0.97	0.11	1.01	0.20	0.93	-0.60	0.38	0.38	77.9	77.6
Lec_2_1	451	308	0.683	0.488	0.356	-0.24	0.11	1.05	1.10	1.07	0.80	0.35	0.39	71.6	71.9
Lec_2_2	451	311	0.690	0.514	0.416	-0.28	0.11	0.99	-0.30	1.02	0.20	0.40	0.39	73.1	72.3
Lec_2_3	451	263	0.583	0.621	0.472	0.28	0.11	0.96	-0.90	0.96	-0.60	0.46	0.43	69.5	68.3
Lec_2_4	451	291	0.647	0.667	0.492	-0.04	0.11	0.92	-1.80	0.91	-1.20	0.47	0.41	73.4	70.1
Lec_2_5	451	314	0.696	0.514	0.377	-0.32	0.11	1.03	0.60	1.01	0.20	0.37	0.39	71.6	72.7
Lec_2_6	451	284	0.631	0.373	0.261	0.04	0.11	1.17	3.90	1.26	3.40	0.27	0.41	62.8	69.4
Lec_2_7	451	277	0.616	0.373	0.287	0.12	0.11	1.15	3.50	1.27	3.60	0.29	0.42	62.3	68.8
Lec_2_8	451	115	0.256	0.480	0.424	2.07	0.12	0.95	-0.80	1.04	0.40	0.48	0.45	79.9	78.7
Lec_2_9	451	237	0.525	0.568	0.419	0.57	0.11	1.02	0.50	1.09	1.60	0.41	0.44	68.8	68.0
Lec_2_10	451	200	0.443	0.532	0.387	0.99	0.11	1.07	1.60	1.08	1.40	0.40	0.46	65.2	69.5
Lec_2_11	451	349	0.774	0.426	0.374	-0.79	0.12	1.01	0.10	0.89	-0.90	0.36	0.35	77.2	78.3

Lec_2_12	451	305	0.676	0.621	0.475	-0.21	0.11	0.93	-1.50	0.88	-1.50	0.46	0.40	73.1	71.4
Lec_2_13	451	329	0.729	0.497	0.417	-0.51	0.12	0.98	-0.40	0.88	-1.20	0.40	0.37	74.5	74.8
Lec_2_14	451	341	0.759	0.641	0.524	-0.67	0.12	0.85	-2.70	0.76	-2.30	0.48	0.36	81.3	76.9
Lec_2_15	451	364	0.807	0.550	0.505	-1.02	0.13	0.85	-2.20	0.68	-2.60	0.46	0.33	83.5	81.1
Exam_1	546	237	0.433	0.307	0.472	0.82	0.10	1.01	0.10	1.00	0.00	0.48	0.48	70.4	69.9
Exam_2	546	383	0.700	0.183	0.571	-0.63	0.11	0.86	-3.10	0.80	-2.70	0.55	0.45	78.5	74.3
Exam_3	546	390	0.713	0.095	0.430	-0.70	0.11	1.04	0.80	1.03	0.40	0.42	0.44	74.0	75.2
Exam_4	546	360	0.658	0.256	0.606	-0.38	0.10	0.83	-4.10	0.72	-4.30	0.59	0.46	76.7	72.0
Exam_5	546	341	0.623	0.146	0.353	-0.18	0.10	1.15	3.60	1.27	3.90	0.35	0.47	65.0	70.7
Exam_6	546	383	0.700	0.132	0.439	-0.63	0.11	1.02	0.40	1.07	0.90	0.43	0.45	75.8	74.3
Exam_7	546	410	0.750	0.154	0.457	-0.94	0.11	0.99	-0.20	0.94	-0.50	0.45	0.43	77.9	77.6
Exam_8	546	268	0.490	0.234	0.420	0.52	0.10	1.08	2.10	1.12	2.00	0.42	0.48	66.0	68.8
Exam_9	546	198	0.362	0.351	0.418	1.22	0.10	1.04	1.00	1.19	2.40	0.43	0.48	73.3	72.5
Exam_10	546	229	0.419	0.256	0.522	0.90	0.10	0.95	-1.40	0.89	-1.70	0.52	0.48	72.7	70.1
Red_Exam_1	546	261	0.477	0.651	0.445	0.43	0.10	0.98	-0.60	1.09	1.90	0.42	0.41	70.3	66.4
Red_Exam_2	546	391	0.715	0.534	0.419	-0.83	0.11	1.05	0.90	1.13	1.50	0.42	0.46	74.1	76.0
Red_Exam_3	546	343	0.627	0.673	0.466	-0.33	0.10	0.98	-0.40	0.98	-0.30	0.45	0.44	71.8	70.1
Red_Exam_4	546	275	0.503	0.592	0.420	0.31	0.10	1.01	0.40	1.02	0.40	0.41	0.41	66.1	66.3
Red_Exam_5	546	410	0.750	0.578	0.470	-1.06	0.11	0.98	-0.40	1.08	0.80	0.47	0.46	79.4	79.0
Red_Exam_6	546	339	0.620	0.680	0.496	-0.29	0.10	0.95	-1.30	0.88	-1.90	0.48	0.44	71.0	69.7
Red_Exam_7	546	182	0.333	0.585	0.400	1.17	0.10	0.98	-0.40	1.00	0.10	0.38	0.37	71.6	71.2
Red_Exam_8	546	290	0.530	0.570	0.379	0.17	0.10	1.07	2.20	1.12	2.30	0.37	0.42	62.9	66.7
Red_Exam_9	546	152	0.278	0.453	0.330	1.48	0.10	1.04	0.80	1.11	1.40	0.32	0.35	73.9	74.5
Red_Exam_10	546	409	0.748	0.673	0.556	-1.05	0.11	0.90	-1.70	0.75	-2.60	0.54	0.46	79.2	78.8
Red_Exam_11	546	399	0.729	0.490	0.413	-0.92	0.11	1.05	0.90	1.11	1.20	0.42	0.46	75.2	77.2
Red_Exam_12	546	277	0.506	0.768	0.521	0.29	0.10	0.90	-3.20	0.88	-2.50	0.48	0.41	70.3	66.3
Red_Exam_13	546	350	0.640	0.790	0.563	-0.39	0.10	0.87	-3.30	0.81	-3.00	0.53	0.44	76.4	70.8
Red_Exam_14	546	420	0.768	0.402	0.395	-1.19	0.12	1.05	0.70	1.44	3.50	0.41	0.47	81.7	80.7
Red_Exam_15	546	448	0.819	0.505	0.504	-1.61	0.13	0.95	-0.60	0.87	-0.90	0.52	0.49	86.5	85.7
Red_Exam_16	546	326	0.596	0.739	0.526	-0.16	0.10	0.91	-2.50	0.85	-2.70	0.50	0.43	72.6	68.6

Red_Exam_17	546	343	0.627	0.548	0.433	-0.33	0.10	1.02	0.60	1.00	0.00	0.43	0.44	68.0	70.1
Red_Exam_18	546	77	0.141	0.183	0.188	2.48	0.13	1.10	1.20	1.37	2.40	0.19	0.28	85.5	85.7
Red_Exam_19	546	315	0.576	0.519	0.367	-0.06	0.10	1.09	2.50	1.14	2.40	0.36	0.43	64.8	67.9
Red_Exam_20	546	315	0.576	0.556	0.413	-0.06	0.10	1.04	1.10	1.03	0.50	0.40	0.43	66.7	67.9
Red_Exam_21	546	344	0.629	0.768	0.553	-0.34	0.10	0.89	-3.00	0.81	-3.00	0.52	0.44	73.5	70.2
Red_Exam_22	546	218	0.399	0.271	0.215	0.83	0.10	1.23	5.90	1.33	5.70	0.23	0.39	59.4	68.2
Red_Exam_23	546	198	0.362	0.548	0.399	1.02	0.10	1.00	0.00	0.99	-0.20	0.38	0.38	68.6	69.8
Red_Exam_24	546	298	0.545	0.607	0.434	0.10	0.10	1.00	0.00	1.04	0.80	0.42	0.42	68.2	67.0
Red_Exam_25	546	334	0.611	0.585	0.454	-0.24	0.10	0.99	-0.20	1.01	0.10	0.44	0.43	69.7	69.3
Red_Exam_26	546	395	0.722	0.534	0.455	-0.88	0.11	1.00	0.00	0.98	-0.20	0.46	0.46	77.9	76.6
Red_Exam_27	546	147	0.269	0.439	0.363	1.54	0.11	0.99	-0.20	1.02	0.30	0.35	0.35	76.8	75.1
Red_Exam_28	546	389	0.711	0.600	0.463	-0.81	0.11	1.00	-0.10	0.98	-0.20	0.46	0.45	77.5	75.7
Red_Exam_29	546	379	0.693	0.753	0.578	-0.70	0.10	0.86	-2.90	0.78	-2.90	0.55	0.45	77.1	74.3
Red_Exam_30	546	159	0.291	0.483	0.357	1.41	0.10	1.01	0.20	1.11	1.50	0.34	0.36	73.7	73.6

Table 53: Breakdown of Individual Items used in Chemistry IB 2012 Multiple-Choice Assessments using CTT and Rasch Analysis

Chemistry IB 2012															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	382	307	0.805	0.323	0.333	-1.17	0.14	1.00	0.10	0.88	-0.80	0.32	0.31	80.5	80.3
Lec_1_2	382	293	0.768	0.271	0.324	-0.93	0.13	1.02	0.30	1.03	0.30	0.31	0.33	76.5	77.1
Lec_1_3	382	231	0.604	0.510	0.460	-0.03	0.11	0.95	-1.20	0.91	-1.30	0.45	0.40	70.1	67.7
Lec_1_4	382	143	0.372	0.542	0.465	1.13	0.12	0.98	-0.40	0.97	-0.40	0.47	0.46	72.5	71.8
Lec_1_5	382	204	0.534	0.469	0.426	0.32	0.11	1.00	0.10	0.96	-0.60	0.43	0.43	66.3	67.2
Lec_1_6	382	216	0.568	0.615	0.515	0.17	0.11	0.90	-2.50	0.87	-2.20	0.50	0.42	70.1	67.0
Lec_1_7	382	206	0.542	0.396	0.356	0.30	0.11	1.07	1.80	1.06	1.00	0.37	0.42	64.2	67.1
Lec_1_8	382	230	0.604	0.542	0.487	-0.02	0.11	0.92	-1.90	0.90	-1.40	0.47	0.40	70.3	67.7
Lec_1_9	382	189	0.497	0.354	0.265	0.52	0.11	1.19	4.30	1.27	4.30	0.27	0.44	58.8	67.5
Lec_1_10	382	236	0.621	0.501	0.442	-0.09	0.12	0.97	-0.80	0.92	-1.00	0.43	0.40	70.3	67.9
Lec_1_11	382	229	0.601	0.334	0.290	0.00	0.11	1.13	3.00	1.22	2.90	0.29	0.41	60.4	67.6
Lec_1_12	382	180	0.469	0.479	0.394	0.63	0.11	1.04	0.90	1.03	0.60	0.41	0.44	65.5	68.1
Lec_1_13	382	205	0.536	0.510	0.459	0.31	0.11	0.97	-0.60	1.03	0.40	0.44	0.42	69.8	67.2
Lec_1_14	382	290	0.758	0.448	0.474	-0.88	0.13	0.89	-1.90	0.74	-2.30	0.44	0.34	77.3	76.4
Lec_1_15	382	248	0.651	0.490	0.435	-0.26	0.12	0.96	-0.80	0.94	-0.70	0.42	0.39	71.7	69.2
Lec_2_1	364	177	0.486	0.363	0.373	0.01	0.12	1.09	2.00	1.17	2.30	0.37	0.44	62.9	68.0
Lec_2_2	364	106	0.291	0.440	0.429	1.07	0.13	1.02	0.40	1.02	0.20	0.44	0.46	77.0	76.6
Lec_2_3	364	205	0.565	0.639	0.572	-0.37	0.12	0.84	-3.90	0.84	-2.30	0.54	0.43	75.3	67.7
Lec_2_4	364	157	0.431	0.341	0.375	0.29	0.12	1.10	2.10	1.09	1.30	0.38	0.45	64.3	69.2
Lec_2_5	364	226	0.621	0.451	0.458	-0.67	0.12	0.97	-0.70	0.92	-1.00	0.44	0.41	70.8	69.2
Lec_2_6	364	151	0.415	0.341	0.375	0.37	0.12	1.09	1.90	1.12	1.60	0.38	0.45	67.7	69.6
Lec_2_7	364	112	0.308	0.495	0.490	0.97	0.13	0.95	-0.70	0.94	-0.50	0.49	0.46	77.5	75.5
Lec_2_8	364	142	0.390	0.681	0.574	0.50	0.12	0.87	-2.80	0.79	-2.80	0.56	0.45	75.0	70.6
Lec_2_9	364	145	0.398	0.407	0.420	0.46	0.12	1.03	0.70	1.07	1.00	0.42	0.45	69.1	70.1
Lec_2_10	364	269	0.739	0.385	0.442	-1.33	0.13	0.94	-1.00	0.86	-1.10	0.42	0.37	75.6	76.0
Lec_2_11	364	224	0.615	0.418	0.440	-0.64	0.12	0.98	-0.50	1.01	0.20	0.43	0.41	69.7	69.0

Lec_2_12	364	209	0.574	0.374	0.435	-0.43	0.12	1.00	0.10	0.94	-0.70	0.43	0.42	66.3	67.8
Lec_2_13	364	219	0.602	0.352	0.384	-0.57	0.12	1.03	0.70	1.22	2.60	0.38	0.42	68.8	68.6
Lec_2_14	364	122	0.336	0.485	0.485	0.81	0.13	0.96	-0.70	0.94	-0.70	0.49	0.46	73.6	73.5
Lec_2_15	364	212	0.582	0.308	0.348	-0.47	0.12	1.09	2.10	1.10	1.30	0.35	0.42	66.9	67.9
Exam_1	433	160	0.370	0.573	0.407	0.49	0.11	1.04	0.90	1.08	1.10	0.40	0.44	71.5	71.2
Exam_2	**No Correct Answer Given Within the Question**														
Exam_3	433	255	0.588	0.654	0.417	-0.63	0.11	1.03	0.80	1.07	1.00	0.41	0.44	68.2	69.6
Exam_4	433	84	0.194	0.258	0.231	1.59	0.13	1.20	2.40	1.31	2.10	0.23	0.39	80.0	82.1
Exam_5	433	277	0.638	0.820	0.530	-0.90	0.11	0.89	-2.40	0.86	-1.80	0.52	0.44	74.3	71.1
Exam_6	433	207	0.477	0.691	0.492	-0.07	0.11	0.95	-1.10	0.94	-1.10	0.48	0.45	71.2	68.4
Exam_7	433	332	0.765	0.682	0.457	-1.66	0.13	0.96	-0.60	0.85	-1.30	0.45	0.41	78.5	79.3
Exam_8	433	120	0.276	0.608	0.474	1.02	0.12	0.93	-1.30	0.93	-0.70	0.47	0.42	78.1	76.0
Exam_9	433	149	0.343	0.590	0.443	0.63	0.11	0.99	-0.10	0.95	-0.60	0.44	0.43	71.7	72.3
Exam_10	433	243	0.560	0.645	0.435	-0.48	0.11	1.02	0.60	1.02	0.30	0.43	0.45	68.6	68.9
Red_Exam_1	421	331	0.764	0.388	0.451	-1.08	0.13	0.98	-0.30	0.86	-1.10	0.35	0.31	78.1	79.2
Red_Exam_2	421	337	0.778	0.240	0.331	-1.18	0.13	1.07	1.00	1.16	1.20	0.22	0.30	80.3	80.4
Red_Exam_3	421	276	0.637	0.462	0.431	-0.33	0.11	1.00	-0.10	0.98	-0.20	0.37	0.36	69.6	69.9
Red_Exam_4	421	154	0.356	0.434	0.382	1.09	0.11	1.01	0.10	1.08	1.20	0.39	0.40	70.8	70.6
Red_Exam_5	421	256	0.591	0.536	0.445	-0.09	0.11	0.98	-0.40	0.99	-0.10	0.39	0.38	68.9	67.9
Red_Exam_6	421	250	0.577	0.674	0.544	-0.02	0.11	0.87	-3.40	0.82	-2.90	0.50	0.38	75.5	67.5
Red_Exam_7	421	237	0.547	0.453	0.387	0.13	0.11	1.04	1.20	1.04	0.70	0.34	0.39	65.1	66.8
Red_Exam_8	421	294	0.679	0.416	0.425	-0.56	0.11	1.00	-0.10	1.06	0.70	0.35	0.35	73.4	72.4
Red_Exam_9	421	218	0.503	0.416	0.360	0.34	0.11	1.07	2.00	1.09	1.60	0.32	0.39	63.2	66.5
Red_Exam_10	421	278	0.642	0.425	0.421	-0.36	0.11	1.00	0.10	1.03	0.40	0.35	0.36	72.0	70.2
Red_Exam_11	421	235	0.543	0.425	0.369	0.15	0.11	1.06	1.70	1.15	2.40	0.32	0.39	65.6	66.8
Red_Exam_12	421	190	0.439	0.397	0.335	0.66	0.11	1.10	2.40	1.13	2.20	0.31	0.40	63.2	67.5
Red_Exam_13	421	218	0.503	0.434	0.390	0.34	0.11	1.04	1.20	1.05	0.80	0.36	0.39	65.1	66.5
Red_Exam_14	421	343	0.792	0.388	0.495	-1.28	0.13	0.92	-1.10	0.77	-1.70	0.39	0.29	83.6	81.7
Red_Exam_15	421	327	0.755	0.296	0.404	-1.02	0.12	1.00	0.10	1.11	0.90	0.30	0.32	80.0	78.4
Red_Exam_16	421	189	0.436	0.545	0.456	0.67	0.11	0.96	-1.10	0.94	-1.00	0.44	0.40	70.1	67.6

Red_Exam_17	421	252	0.582	0.785	0.623	-0.04	0.11	0.79	-5.70	0.73	-4.50	0.58	0.38	79.8	67.6
Red_Exam_18	421	183	0.423	0.462	0.401	0.74	0.11	1.01	0.30	0.99	-0.10	0.40	0.40	66.3	67.9
Red_Exam_19	421	280	0.647	0.453	0.436	-0.38	0.11	0.99	-0.30	1.02	0.30	0.37	0.36	70.5	70.4
Red_Exam_20	421	183	0.423	0.462	0.403	0.74	0.11	1.02	0.50	1.00	0.00	0.39	0.40	65.3	67.9
Red_Exam_21	421	225	0.520	0.443	0.359	0.26	0.11	1.08	2.10	1.14	2.40	0.31	0.39	61.8	66.6
Red_Exam_22	421	194	0.448	0.600	0.500	0.62	0.11	0.91	-2.40	0.89	-2.10	0.49	0.40	72.4	67.2
Red_Exam_23	421	146	0.338	0.454	0.394	1.19	0.11	1.00	0.10	1.02	0.40	0.40	0.40	72.0	71.6
Red_Exam_24	421	157	0.363	0.583	0.431	1.05	0.11	0.96	-0.80	0.99	-0.10	0.43	0.40	75.8	70.3
Red_Exam_25	421	361	0.834	0.342	0.490	-1.63	0.15	0.92	-0.90	0.74	-1.60	0.36	0.27	86.2	85.8
Red_Exam_26	421	260	0.600	0.425	0.427	-0.14	0.11	1.00	0.10	1.05	0.70	0.37	0.37	67.0	68.2
Red_Exam_27	421	298	0.688	0.351	0.361	-0.61	0.11	1.06	1.10	1.09	0.90	0.28	0.35	72.9	73.1
Red_Exam_28	421	178	0.411	0.619	0.497	0.80	0.11	0.91	-2.30	0.89	-2.00	0.49	0.40	74.3	68.3
Red_Exam_29	421	276	0.637	0.333	0.333	-0.33	0.11	1.09	2.10	1.29	3.40	0.25	0.36	67.2	69.9
Red_Exam_30	421	224	0.517	0.462	0.391	0.28	0.11	1.04	1.10	1.07	1.20	0.35	0.39	65.8	66.5

Table 54: Breakdown of Individual Items used in Chemistry IB 2013 Multiple-Choice Assessments using CTT and Rasch Analysis

Chemistry IB 2013															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	378	293	0.775	0.222	0.386	-0.87	0.13	0.99	-0.10	0.85	-1.30	0.37	0.35	77.3	77.8
Lec_1_2	378	288	0.762	0.296	0.400	-0.78	0.13	0.97	-0.40	0.99	0.00	0.37	0.35	75.4	76.7
Lec_1_3	378	257	0.680	0.413	0.484	-0.30	0.12	0.91	-1.80	0.96	-0.50	0.45	0.39	77.6	71.3
Lec_1_4	378	142	0.376	0.423	0.419	1.26	0.12	1.04	0.70	1.05	0.70	0.43	0.46	69.4	71.3
Lec_1_5	378	211	0.558	0.519	0.468	0.33	0.12	0.97	-0.80	0.94	-1.10	0.46	0.43	69.9	67.5
Lec_1_6	378	236	0.624	0.423	0.437	-0.01	0.12	0.98	-0.40	1.01	0.10	0.42	0.41	71.9	68.9
Lec_1_7	378	214	0.566	0.444	0.393	0.29	0.12	1.04	0.90	1.01	0.20	0.40	0.43	62.6	67.6
Lec_1_8	378	243	0.643	0.392	0.421	-0.10	0.12	1.00	-0.10	0.98	-0.30	0.41	0.40	68.9	69.5
Lec_1_9	378	202	0.534	0.370	0.367	0.45	0.11	1.08	1.90	1.09	1.60	0.37	0.44	62.3	67.2
Lec_1_10	378	248	0.658	0.340	0.410	-0.17	0.12	1.00	0.00	1.00	0.10	0.40	0.40	68.6	70.0
Lec_1_11	378	238	0.630	0.286	0.336	-0.03	0.12	1.09	1.90	1.08	1.20	0.34	0.41	65.8	69.1
Lec_1_12	378	173	0.458	0.307	0.312	0.83	0.12	1.13	2.80	1.21	3.30	0.34	0.45	64.8	68.3
Lec_1_13	378	225	0.595	0.392	0.415	0.14	0.12	1.01	0.30	1.00	0.00	0.41	0.42	67.2	68.0
Lec_1_14	378	299	0.793	0.361	0.521	-0.98	0.14	0.84	-2.30	0.69	-2.60	0.48	0.34	82.8	79.1
Lec_1_15	378	240	0.635	0.466	0.476	-0.06	0.12	0.94	-1.40	0.94	-0.80	0.46	0.41	73.0	69.3
Lec_2_1	348	183	0.527	0.761	0.422	0.00	0.12	1.02	0.40	1.05	0.70	0.42	0.43	67.1	68.5
Lec_2_2	348	107	0.307	0.448	0.338	1.13	0.13	1.09	1.40	1.20	1.90	0.35	0.43	70.8	74.4
Lec_2_3	348	252	0.724	0.816	0.498	-1.05	0.13	0.89	-1.90	0.76	-2.30	0.49	0.38	78.0	75.1
Lec_2_4	348	187	0.539	0.634	0.382	-0.06	0.12	1.06	1.30	1.05	0.70	0.38	0.43	65.9	68.5
Lec_2_5	348	246	0.707	0.782	0.433	-0.95	0.13	0.96	-0.60	0.97	-0.20	0.43	0.39	72.8	73.9
Lec_2_6	348	140	0.402	0.621	0.368	0.62	0.12	1.08	1.50	1.19	2.40	0.36	0.44	68.8	70.4
Lec_2_7	348	114	0.328	0.471	0.298	1.02	0.13	1.15	2.60	1.18	1.90	0.31	0.43	67.6	73.5
Lec_2_8	348	154	0.443	0.770	0.468	0.41	0.12	0.96	-0.90	1.00	0.00	0.46	0.44	72.0	69.2
Lec_2_9	348	137	0.395	0.749	0.503	0.66	0.12	0.92	-1.50	0.92	-1.10	0.50	0.44	72.5	70.7
Lec_2_10	348	275	0.790	0.678	0.406	-1.47	0.14	0.97	-0.40	0.85	-1.00	0.39	0.35	79.8	80.2
Lec_2_11	348	210	0.603	0.805	0.485	-0.39	0.12	0.93	-1.60	0.91	-1.20	0.48	0.42	74.6	69.1

Lec_2_12	348	216	0.621	0.793	0.475	-0.48	0.12	0.95	-1.10	0.87	-1.60	0.47	0.42	69.1	69.4
Lec_2_13	348	182	0.524	0.519	0.310	0.01	0.12	1.14	3.10	1.14	2.00	0.32	0.43	62.7	68.5
Lec_2_14	348	131	0.378	0.888	0.566	0.75	0.12	0.85	-3.00	0.82	-2.40	0.56	0.44	79.5	71.2
Lec_2_15	348	198	0.571	0.738	0.413	-0.22	0.12	1.02	0.40	1.09	1.30	0.40	0.43	67.1	68.6
Exam_1	450	186	0.414	0.570	0.453	0.57	0.11	1.00	-0.10	1.07	1.00	0.45	0.46	71.0	70.4
Exam_2	450	343	0.762	0.169	0.341	-1.36	0.12	1.09	1.40	1.28	2.10	0.34	0.42	77.8	79.4
Exam_3	450	316	0.702	0.516	0.548	-0.97	0.12	0.87	-2.60	0.79	-2.20	0.54	0.44	80.0	75.4
Exam_4	450	110	0.245	0.249	0.277	1.55	0.12	1.15	2.30	1.49	3.50	0.29	0.42	75.7	78.8
Exam_5	450	323	0.718	0.373	0.486	-1.07	0.12	0.93	-1.10	0.87	-1.20	0.48	0.43	81.1	76.2
Exam_6	450	232	0.516	0.596	0.511	0.04	0.11	0.94	-1.40	0.92	-1.20	0.51	0.46	73.7	69.2
Exam_7	450	320	0.711	0.364	0.436	-1.03	0.12	1.00	0.00	1.02	0.30	0.44	0.44	75.5	75.8
Exam_8	450	99	0.220	0.436	0.386	1.72	0.13	1.02	0.30	1.16	1.20	0.39	0.41	81.1	80.3
Exam_9	450	165	0.367	0.676	0.510	0.82	0.11	0.92	-1.70	0.94	-0.70	0.50	0.45	76.6	72.1
Exam_10	450	259	0.576	0.524	0.450	-0.27	0.11	1.02	0.40	1.02	0.30	0.44	0.46	70.6	70.3
Red_Exam_1	434	366	0.813	0.320	0.447	-1.38	0.14	1.01	0.10	1.09	0.60	0.29	0.30	84.9	84.4
Red_Exam_2	434	333	0.740	0.391	0.417	-0.82	0.12	1.07	1.10	1.06	0.50	0.30	0.35	74.4	77.7
Red_Exam_3	434	292	0.649	0.391	0.426	-0.27	0.11	1.05	1.20	1.04	0.50	0.35	0.39	67.0	71.3
Red_Exam_4	434	155	0.344	0.507	0.391	1.35	0.11	1.06	1.20	1.09	1.20	0.39	0.45	72.3	72.6
Red_Exam_5	434	284	0.631	0.533	0.506	-0.17	0.11	0.96	-1.00	0.90	-1.20	0.44	0.40	71.2	70.4
Red_Exam_6	434	296	0.658	0.551	0.494	-0.32	0.11	0.97	-0.60	0.90	-1.20	0.42	0.39	73.0	71.8
Red_Exam_7	434	296	0.658	0.498	0.470	-0.32	0.11	1.00	0.00	0.96	-0.40	0.39	0.39	74.0	71.8
Red_Exam_8	434	315	0.700	0.507	0.483	-0.57	0.12	0.98	-0.40	0.93	-0.70	0.39	0.37	75.6	74.5
Red_Exam_9	434	258	0.575	0.383	0.333	0.14	0.11	1.17	4.10	1.29	3.90	0.26	0.42	63.3	68.5
Red_Exam_10	434	303	0.673	0.542	0.532	-0.41	0.11	0.92	-1.70	0.90	-1.10	0.45	0.38	75.1	72.7
Red_Exam_11	434	266	0.591	0.471	0.423	0.05	0.11	1.06	1.40	1.05	0.70	0.37	0.41	66.0	68.9
Red_Exam_12	434	229	0.509	0.409	0.365	0.47	0.11	1.14	3.30	1.15	2.40	0.32	0.43	61.2	68.4
Red_Exam_13	434	249	0.553	0.471	0.414	0.24	0.11	1.07	1.70	1.12	1.80	0.36	0.42	66.3	68.3
Red_Exam_14	434	347	0.771	0.507	0.568	-1.04	0.13	0.87	-1.90	0.74	-2.10	0.45	0.33	81.4	80.4
Red_Exam_15	434	317	0.704	0.524	0.493	-0.60	0.12	0.96	-0.80	0.97	-0.30	0.39	0.37	77.4	74.8
Red_Exam_16	434	227	0.504	0.587	0.471	0.50	0.11	0.99	-0.30	1.00	0.10	0.44	0.43	70.5	68.4

Red_Exam_17	434	278	0.618	0.684	0.608	-0.10	0.11	0.83	-4.20	0.74	-3.80	0.55	0.40	75.6	69.8
Red_Exam_18	434	244	0.542	0.533	0.460	0.30	0.11	1.01	0.30	1.00	0.00	0.42	0.43	65.6	68.3
Red_Exam_19	434	301	0.669	0.516	0.495	-0.38	0.11	0.97	-0.60	0.95	-0.50	0.41	0.38	72.3	72.5
Red_Exam_20	434	201	0.447	0.480	0.431	0.80	0.11	1.04	1.00	1.10	1.70	0.40	0.44	68.4	69.2
Red_Exam_21	434	243	0.540	0.373	0.341	0.31	0.11	1.17	4.00	1.23	3.50	0.28	0.43	61.2	68.3
Red_Exam_22	434	196	0.436	0.640	0.550	0.85	0.11	0.87	-3.20	0.88	-2.10	0.54	0.44	78.4	69.5
Red_Exam_23	434	155	0.344	0.524	0.441	1.35	0.11	1.01	0.20	1.01	0.20	0.44	0.45	72.8	72.6
Red_Exam_24	434	174	0.387	0.604	0.471	1.11	0.11	0.96	-1.00	0.99	-0.20	0.48	0.45	72.3	70.9
Red_Exam_25	434	356	0.791	0.418	0.505	-1.20	0.13	0.95	-0.70	0.85	-1.00	0.37	0.32	82.1	82.3
Red_Exam_26	434	309	0.687	0.524	0.497	-0.49	0.12	0.96	-0.70	0.94	-0.60	0.41	0.38	73.7	73.6
Red_Exam_27	434	290	0.644	0.516	0.471	-0.24	0.11	1.00	0.00	1.05	0.60	0.39	0.39	71.2	71.1
Red_Exam_28	434	202	0.450	0.650	0.548	0.78	0.11	0.88	-2.90	0.86	-2.40	0.54	0.44	74.7	69.1
Red_Exam_29	434	284	0.631	0.507	0.452	-0.17	0.11	1.02	0.50	1.10	1.20	0.38	0.40	70.2	70.4
Red_Exam_30	434	251	0.559	0.543	0.480	0.22	0.11	0.99	-0.20	0.95	-0.80	0.44	0.42	67.7	68.4

Table 55: Breakdown of Individual Items used in Chemistry IB 2014 Multiple-Choice Assessments using CTT and Rasch Analysis

Chemistry IB 2014															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	423	362	0.856	0.312	0.395	-1.52	0.15	0.94	-0.70	0.75	-1.50	0.38	0.32	85.9	85.2
Lec_1_2	423	326	0.771	0.473	0.450	-0.85	0.13	0.94	-0.90	0.97	-0.20	0.42	0.38	77.9	77.7
Lec_1_3	423	276	0.656	0.570	0.429	-0.14	0.11	1.03	0.60	1.03	0.40	0.42	0.44	70.0	70.6
Lec_1_4	423	189	0.448	0.607	0.475	0.96	0.11	1.02	0.50	1.01	0.10	0.49	0.50	69.0	70.6
Lec_1_5	423	247	0.584	0.690	0.502	0.23	0.11	0.96	-0.90	0.93	-1.00	0.50	0.46	70.7	69.1
Lec_1_6	423	274	0.649	0.730	0.542	-0.11	0.11	0.90	-2.30	0.82	-2.30	0.52	0.44	75.2	70.4
Lec_1_7	423	280	0.662	0.596	0.473	-0.19	0.12	0.97	-0.60	0.94	-0.70	0.46	0.44	71.5	71.0
Lec_1_8	423	267	0.631	0.615	0.446	-0.02	0.11	1.02	0.40	0.98	-0.20	0.44	0.45	69.0	70.1
Lec_1_9	423	209	0.494	0.520	0.390	0.71	0.11	1.13	2.80	1.15	2.20	0.40	0.49	63.8	69.3
Lec_1_10	423	253	0.600	0.749	0.566	0.16	0.11	0.88	-2.80	0.88	-1.80	0.54	0.46	74.2	69.2
Lec_1_11	423	264	0.626	0.474	0.388	0.02	0.11	1.09	1.90	1.15	1.90	0.38	0.45	67.2	69.9
Lec_1_12	423	193	0.456	0.340	0.262	0.91	0.11	1.28	5.60	1.45	6.10	0.29	0.50	60.0	70.3
Lec_1_13	423	250	0.591	0.700	0.515	0.20	0.11	0.95	-1.20	0.91	-1.40	0.50	0.46	68.5	69.2
Lec_1_14	423	302	0.714	0.577	0.475	-0.49	0.12	0.95	-1.00	0.88	-1.20	0.46	0.41	74.9	73.8
Lec_1_15	423	255	0.603	0.671	0.516	0.13	0.11	0.94	-1.40	0.92	-1.20	0.50	0.46	70.5	69.3
Lec_2_1	394	215	0.547	0.537	0.469	-0.10	0.11	0.97	-0.60	0.97	-0.50	0.46	0.44	70.3	68.3
Lec_2_2	394	130	0.329	0.476	0.404	1.03	0.12	1.06	1.00	1.07	0.80	0.42	0.46	73.9	74.3
Lec_2_3	394	277	0.704	0.719	0.581	-0.92	0.12	0.79	-4.30	0.69	-3.40	0.55	0.38	79.3	73.0
Lec_2_4	394	200	0.506	0.486	0.422	0.10	0.11	1.03	0.70	1.07	1.00	0.42	0.44	68.0	68.4
Lec_2_5	394	269	0.684	0.476	0.456	-0.81	0.12	0.94	-1.30	0.90	-1.00	0.44	0.39	73.1	71.7
Lec_2_6	394	159	0.403	0.496	0.441	0.63	0.12	1.02	0.50	0.99	-0.10	0.45	0.46	66.7	71.0
Lec_2_7	394	127	0.322	0.435	0.414	1.08	0.12	1.04	0.70	1.07	0.80	0.43	0.46	72.6	74.7
Lec_2_8	394	169	0.430	0.648	0.516	0.49	0.11	0.94	-1.40	0.94	-0.90	0.50	0.46	72.4	70.0
Lec_2_9	394	142	0.363	0.487	0.454	0.86	0.12	1.00	-0.10	1.08	1.00	0.45	0.46	74.4	72.8
Lec_2_10	394	311	0.790	0.324	0.373	-1.46	0.13	0.97	-0.50	0.94	-0.40	0.36	0.34	80.6	79.4
Lec_2_11	394	243	0.618	0.385	0.349	-0.46	0.11	1.08	1.70	1.13	1.70	0.35	0.42	69.3	69.0

Lec_2_12	394	238	0.605	0.415	0.422	-0.39	0.11	1.02	0.50	0.96	-0.60	0.41	0.42	66.7	68.8
Lec_2_13	394	224	0.570	0.405	0.357	-0.21	0.11	1.09	2.20	1.11	1.60	0.36	0.43	64.6	68.5
Lec_2_14	394	159	0.405	0.587	0.494	0.63	0.12	0.97	-0.60	0.98	-0.30	0.48	0.46	72.4	71.0
Lec_2_15	394	244	0.620	0.425	0.391	-0.47	0.11	1.04	0.90	1.10	1.20	0.38	0.41	68.0	69.1
Exam_1	486	181	0.372	0.568	0.395	0.79	0.11	1.10	2.00	1.16	2.10	0.40	0.47	69.0	72.4
Exam_2	486	348	0.716	0.626	0.411	-1.06	0.11	1.02	0.40	1.09	0.90	0.41	0.43	74.3	75.2
Exam_3	486	305	0.630	0.727	0.466	-0.55	0.11	0.99	-0.10	0.91	-1.20	0.46	0.45	69.6	70.8
Exam_4	486	133	0.274	0.453	0.372	1.38	0.12	1.12	2.00	1.16	1.50	0.37	0.46	74.9	77.9
Exam_5	486	342	0.704	0.757	0.503	-0.98	0.11	0.91	-1.70	0.86	-1.40	0.49	0.43	78.7	74.6
Exam_6	486	261	0.539	0.760	0.511	-0.07	0.10	0.95	-1.30	0.88	-1.90	0.51	0.46	69.4	68.7
Exam_7	486	367	0.755	0.584	0.373	-1.31	0.12	1.06	1.00	1.10	0.90	0.37	0.41	77.4	78.3
Exam_8	486	119	0.245	0.528	0.461	1.58	0.12	0.98	-0.30	0.95	-0.40	0.47	0.45	80.4	79.7
Exam_9	486	199	0.410	0.693	0.499	0.59	0.11	0.97	-0.70	0.92	-1.10	0.50	0.47	71.1	70.8
Exam_10	486	288	0.593	0.856	0.518	-0.36	0.10	0.93	-1.70	0.89	-1.50	0.51	0.46	71.7	70.0
Red_Exam_1	456	397	0.819	0.462	0.574	-1.65	0.15	0.93	-0.70	0.84	-0.90	0.33	0.27	87.6	87.1
Red_Exam_2	456	353	0.728	0.470	0.486	-0.89	0.12	1.02	0.40	1.07	0.70	0.30	0.33	77.8	78.2
Red_Exam_3	456	295	0.608	0.627	0.524	-0.16	0.11	0.95	-1.30	1.06	0.90	0.41	0.38	73.6	69.3
Red_Exam_4	456	183	0.377	0.553	0.464	1.05	0.11	0.97	-0.60	0.94	-0.90	0.45	0.42	70.2	69.8
Red_Exam_5	456	299	0.616	0.643	0.565	-0.21	0.11	0.90	-2.40	0.84	-2.30	0.46	0.37	72.7	69.8
Red_Exam_6	456	290	0.598	0.602	0.518	-0.11	0.11	0.96	-1.00	0.90	-1.40	0.42	0.38	68.7	68.8
Red_Exam_7	456	301	0.621	0.404	0.421	-0.23	0.11	1.09	2.00	1.11	1.50	0.30	0.37	65.1	70.0
Red_Exam_8	456	323	0.666	0.445	0.452	-0.49	0.11	1.05	1.00	1.13	1.50	0.30	0.36	71.1	73.0
Red_Exam_9	456	258	0.532	0.487	0.404	0.24	0.10	1.09	2.60	1.10	1.60	0.32	0.40	60.2	66.7
Red_Exam_10	456	309	0.637	0.685	0.570	-0.32	0.11	0.90	-2.30	0.82	-2.40	0.46	0.37	73.1	71.0
Red_Exam_11	456	266	0.548	0.313	0.306	0.16	0.10	1.22	5.70	1.32	4.70	0.19	0.39	57.6	67.0
Red_Exam_12	456	259	0.534	0.487	0.401	0.23	0.10	1.10	2.70	1.17	2.80	0.30	0.40	60.0	66.8
Red_Exam_13	456	258	0.533	0.488	0.417	0.24	0.10	1.08	2.10	1.16	2.50	0.32	0.40	64.2	66.7
Red_Exam_14	456	351	0.724	0.586	0.571	-0.86	0.12	0.90	-1.60	0.83	-1.60	0.41	0.33	80.4	77.8
Red_Exam_15	456	311	0.641	0.561	0.523	-0.35	0.11	0.96	-1.00	0.93	-0.80	0.40	0.37	73.6	71.3
Red_Exam_16	456	217	0.447	0.586	0.475	0.68	0.10	0.97	-0.90	0.99	-0.20	0.44	0.42	70.4	67.3

Red_Exam_17	456	301	0.621	0.767	0.635	-0.23	0.11	0.81	-4.80	0.72	-4.20	0.54	0.37	77.1	70.0
Red_Exam_18	456	233	0.480	0.569	0.451	0.51	0.10	1.01	0.30	1.01	0.20	0.40	0.41	67.1	66.7
Red_Exam_19	456	295	0.608	0.544	0.485	-0.16	0.11	1.00	0.10	1.02	0.30	0.37	0.38	69.6	69.3
Red_Exam_20	456	208	0.429	0.437	0.380	0.77	0.10	1.10	2.60	1.10	1.70	0.33	0.42	62.7	67.8
Red_Exam_21	456	284	0.586	0.404	0.389	-0.04	0.11	1.12	3.00	1.23	3.20	0.27	0.38	62.4	68.2
Red_Exam_22	456	220	0.455	0.736	0.565	0.65	0.10	0.86	-4.00	0.81	-3.60	0.54	0.41	73.3	67.2
Red_Exam_23	456	161	0.332	0.478	0.399	1.30	0.11	1.04	0.80	1.06	0.90	0.39	0.43	73.3	72.0
Red_Exam_24	456	182	0.375	0.627	0.475	1.06	0.11	0.93	-1.50	0.95	-0.70	0.47	0.42	74.0	69.9
Red_Exam_25	456	376	0.775	0.396	0.523	-1.25	0.13	0.97	-0.30	0.96	-0.20	0.32	0.30	83.8	82.7
Red_Exam_26	456	311	0.641	0.569	0.487	-0.35	0.11	1.01	0.30	0.96	-0.50	0.36	0.37	70.9	71.3
Red_Exam_27	456	311	0.641	0.495	0.455	-0.35	0.11	1.04	1.00	1.06	0.70	0.32	0.37	70.4	71.3
Red_Exam_28	456	217	0.448	0.702	0.555	0.68	0.10	0.86	-3.90	0.83	-3.10	0.53	0.42	76.7	67.3
Red_Exam_29	456	296	0.610	0.536	0.483	-0.17	0.11	1.00	0.00	1.14	1.90	0.36	0.38	71.1	69.4
Red_Exam_30	456	256	0.529	0.579	0.472	0.26	0.10	1.01	0.20	1.00	0.10	0.39	0.40	65.6	66.7

Table 56: Breakdown of Individual Items used in Chemistry IB 2015 Multiple-Choice Assessments using CTT and Rasch Analysis

Chemistry IB 2015															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	429	364	0.848	0.289	0.361	-1.55	0.14	0.96	-0.40	0.79	-1.30	0.34	0.29	84.9	84.6
Lec_1_2	429	331	0.772	0.345	0.375	-0.98	0.12	0.99	-0.20	0.98	-0.10	0.36	0.35	77.2	77.5
Lec_1_3	429	266	0.620	0.494	0.434	-0.11	0.11	0.99	-0.20	0.93	-1.00	0.43	0.42	68.3	68.6
Lec_1_4	429	162	0.378	0.569	0.493	1.13	0.11	0.96	-0.90	0.97	-0.40	0.50	0.47	74.3	72.2
Lec_1_5	429	234	0.545	0.587	0.527	0.27	0.11	0.91	-2.50	0.92	-1.30	0.51	0.44	74.6	67.7
Lec_1_6	429	244	0.569	0.606	0.526	0.15	0.11	0.91	-2.50	0.87	-2.20	0.51	0.43	71.2	67.7
Lec_1_7	429	231	0.538	0.578	0.471	0.30	0.11	0.97	-0.70	0.93	-1.20	0.47	0.44	69.1	67.8
Lec_1_8	429	280	0.653	0.401	0.404	-0.28	0.11	1.01	0.20	1.05	0.70	0.39	0.40	68.6	69.9
Lec_1_9	429	228	0.531	0.373	0.311	0.34	0.11	1.16	3.80	1.25	4.00	0.31	0.44	61.2	67.8
Lec_1_10	429	276	0.645	0.411	0.421	-0.23	0.11	0.99	-0.10	1.00	0.00	0.41	0.41	69.1	69.5
Lec_1_11	429	250	0.585	0.337	0.332	0.08	0.11	1.11	2.80	1.13	2.00	0.34	0.43	64.0	68.0
Lec_1_12	429	182	0.424	0.429	0.338	0.88	0.11	1.14	3.00	1.16	2.50	0.36	0.47	64.7	70.3
Lec_1_13	429	233	0.544	0.589	0.482	0.28	0.11	0.96	-1.00	0.95	-0.80	0.47	0.44	67.6	67.7
Lec_1_14	429	297	0.694	0.542	0.507	-0.50	0.11	0.89	-2.50	0.81	-2.30	0.48	0.39	77.7	71.8
Lec_1_15	429	237	0.552	0.494	0.396	0.23	0.11	1.05	1.30	1.07	1.20	0.40	0.44	63.8	67.7
Lec_2_1	392	187	0.477	0.459	0.413	0.43	0.11	1.02	0.60	1.07	1.10	0.41	0.43	66.4	67.6
Lec_2_2	392	98	0.250	0.276	0.313	1.70	0.13	1.14	1.90	1.25	2.10	0.34	0.45	77.1	79.4
Lec_2_3	392	254	0.646	0.499	0.510	-0.42	0.12	0.89	-2.60	0.83	-2.20	0.48	0.39	74.7	69.5
Lec_2_4	392	196	0.499	0.305	0.356	0.32	0.11	1.09	2.10	1.07	1.20	0.36	0.43	63.5	67.4
Lec_2_5	392	284	0.727	0.276	0.399	-0.85	0.12	0.97	-0.50	1.03	0.30	0.38	0.36	75.5	74.1
Lec_2_6	392	233	0.593	0.305	0.338	-0.15	0.11	1.09	2.10	1.14	2.00	0.33	0.41	65.6	67.7
Lec_2_7	392	352	0.898	0.041	0.303	-2.21	0.17	0.96	-0.30	0.80	-0.90	0.29	0.25	89.8	89.7
Lec_2_8	392	130	0.333	0.499	0.454	1.19	0.12	0.98	-0.30	0.98	-0.20	0.46	0.45	74.7	73.4
Lec_2_9	392	151	0.385	0.622	0.520	0.90	0.12	0.91	-2.00	0.89	-1.60	0.52	0.45	75.5	70.7
Lec_2_10	392	308	0.784	0.254	0.426	-1.23	0.13	0.91	-1.30	0.96	-0.30	0.40	0.33	80.5	79.0
Lec_2_11	392	216	0.550	0.346	0.369	0.06	0.11	1.06	1.60	1.09	1.40	0.36	0.42	65.6	67.3

Lec_2_12	392	246	0.626	0.458	0.458	-0.32	0.11	0.95	-1.10	0.91	-1.10	0.44	0.40	72.1	68.9
Lec_2_13	392	238	0.608	0.316	0.349	-0.21	0.11	1.06	1.40	1.10	1.30	0.35	0.40	65.9	68.2
Lec_2_14	392	155	0.394	0.651	0.532	0.84	0.12	0.89	-2.30	0.87	-1.90	0.53	0.45	76.0	70.1
Lec_2_15	392	224	0.570	0.377	0.405	-0.04	0.11	1.02	0.50	1.02	0.40	0.40	0.41	66.7	67.5
Exam_1	487	187	0.384	0.386	0.321	0.58	0.10	1.16	3.50	1.21	2.90	0.33	0.45	65.1	71.2
Exam_2	487	330	0.678	0.575	0.443	-0.96	0.11	1.00	0.10	0.98	-0.20	0.44	0.44	73.9	73.4
Exam_3	487	279	0.574	0.658	0.490	-0.39	0.10	0.97	-0.80	0.91	-1.50	0.49	0.46	70.1	69.6
Exam_4	487	126	0.259	0.329	0.277	1.31	0.12	1.18	3.00	1.29	2.70	0.28	0.42	72.4	77.6
Exam_5	487	316	0.649	0.715	0.560	-0.80	0.11	0.87	-3.00	0.79	-2.90	0.55	0.45	76.4	72.3
Exam_6	487	244	0.502	0.798	0.577	-0.02	0.10	0.86	-3.60	0.81	-3.40	0.57	0.46	73.7	68.7
Exam_7	487	341	0.700	0.550	0.476	-1.09	0.11	0.95	-0.90	0.87	-1.40	0.48	0.43	76.4	74.8
Exam_8	487	123	0.253	0.501	0.409	1.35	0.12	1.01	0.10	0.99	0.00	0.41	0.42	78.1	78.0
Exam_9	487	194	0.398	0.624	0.476	0.50	0.10	0.96	-0.80	0.99	-0.20	0.47	0.45	72.9	70.5
Exam_10	487	287	0.592	0.553	0.416	-0.48	0.10	1.05	1.20	1.11	1.70	0.41	0.45	68.9	69.7
Red_Exam_1	472	380	0.780	0.526	0.486	-1.07	0.12	0.94	-0.90	0.89	-0.80	0.36	0.31	81.1	80.4
Red_Exam_2	472	374	0.768	0.493	0.440	-0.99	0.12	0.98	-0.30	1.08	0.60	0.32	0.32	80.3	79.3
Red_Exam_3	472	282	0.580	0.658	0.482	0.13	0.10	0.98	-0.50	0.98	-0.30	0.43	0.41	68.7	68.4
Red_Exam_4	472	178	0.366	0.591	0.471	1.25	0.11	0.98	-0.30	0.98	-0.30	0.48	0.47	71.9	72.5
Red_Exam_5	472	258	0.530	0.706	0.532	0.38	0.10	0.92	-2.20	0.94	-0.90	0.49	0.43	73.2	68.2
Red_Exam_6	472	279	0.573	0.764	0.555	0.16	0.10	0.89	-3.00	0.83	-2.60	0.51	0.42	72.3	68.3
Red_Exam_7	472	284	0.583	0.517	0.386	0.11	0.10	1.10	2.50	1.16	2.20	0.33	0.41	64.8	68.5
Red_Exam_8	472	357	0.733	0.534	0.438	-0.75	0.12	1.01	0.20	0.98	-0.10	0.33	0.34	76.4	76.2
Red_Exam_9	472	262	0.538	0.419	0.360	0.34	0.10	1.14	3.50	1.26	3.80	0.31	0.43	64.6	68.1
Red_Exam_10	472	343	0.704	0.575	0.474	-0.57	0.11	0.97	-0.70	0.98	-0.10	0.38	0.36	76.0	74.0
Red_Exam_11	472	297	0.610	0.509	0.381	-0.04	0.10	1.10	2.40	1.15	1.80	0.32	0.40	66.3	69.2
Red_Exam_12	472	273	0.561	0.435	0.325	0.22	0.10	1.18	4.50	1.27	3.80	0.27	0.42	59.2	68.2
Red_Exam_13	472	265	0.544	0.723	0.516	0.31	0.10	0.94	-1.70	0.89	-1.80	0.48	0.43	71.2	68.1
Red_Exam_14	472	324	0.665	0.706	0.536	-0.34	0.11	0.90	-2.30	0.83	-1.90	0.46	0.38	74.9	71.5
Red_Exam_15	472	353	0.725	0.583	0.496	-0.70	0.11	0.94	-1.20	0.89	-1.00	0.39	0.34	79.0	75.5
Red_Exam_16	472	229	0.470	0.674	0.496	0.69	0.10	0.96	-1.10	0.92	-1.40	0.48	0.45	70.2	68.9

Red_Exam_17	472	311	0.641	0.759	0.572	-0.19	0.11	0.87	-3.40	0.81	-2.40	0.50	0.39	74.9	70.2
Red_Exam_18	472	255	0.524	0.517	0.379	0.41	0.10	1.11	2.80	1.10	1.60	0.35	0.43	62.2	68.2
Red_Exam_19	472	326	0.669	0.608	0.499	-0.36	0.11	0.95	-1.10	0.92	-0.80	0.42	0.37	71.5	71.7
Red_Exam_20	472	267	0.548	0.485	0.393	0.29	0.10	1.09	2.40	1.08	1.20	0.35	0.42	64.4	68.1
Red_Exam_21	472	438	0.899	0.353	0.455	-2.31	0.18	0.98	-0.10	0.69	-1.20	0.24	0.20	92.7	92.7
Red_Exam_22	472	215	0.442	0.733	0.512	0.84	0.10	0.94	-1.60	0.90	-1.70	0.50	0.45	71.9	69.6
Red_Exam_23	472	166	0.341	0.501	0.381	1.39	0.11	1.10	2.00	1.18	2.40	0.38	0.47	70.6	73.7
Red_Exam_24	472	184	0.379	0.683	0.499	1.18	0.11	0.92	-1.70	0.94	-1.00	0.52	0.46	75.8	71.9
Red_Exam_25	472	358	0.735	0.550	0.460	-0.76	0.12	0.99	-0.20	0.88	-1.00	0.36	0.34	76.6	76.3
Red_Exam_26	472	296	0.608	0.608	0.461	-0.02	0.10	1.00	0.10	1.04	0.60	0.40	0.40	68.7	69.1
Red_Exam_27	472	326	0.671	0.436	0.336	-0.36	0.11	1.13	2.90	1.34	3.30	0.25	0.37	68.9	71.7
Red_Exam_28	472	224	0.460	0.821	0.568	0.74	0.10	0.87	-3.40	0.81	-3.50	0.56	0.45	74.2	69.1
Red_Exam_29	472	304	0.624	0.542	0.409	-0.11	0.11	1.06	1.50	1.08	1.00	0.34	0.39	69.1	69.6
Red_Exam_30	472	281	0.577	0.665	0.471	0.14	0.10	0.99	-0.20	1.05	0.70	0.41	0.41	69.7	68.3

Table 57: Breakdown of Individual Items used in Foundations of Chemistry IA 2012 Multiple-Choice Assessments using CTT and Rasch Analysis

Foundations of Chemistry IA 2012															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	259	244	0.942	-0.046	0.164	-1.70	0.28	1.08	0.40	1.18	0.60	0.17	0.23	93.3	93.7
Lec_1_2	259	200	0.772	0.340	0.423	0.09	0.16	0.99	-0.10	0.96	-0.20	0.41	0.40	78.6	77.0
Lec_1_3	259	240	0.927	0.046	0.207	-1.42	0.25	1.09	0.50	0.97	0.00	0.21	0.26	91.6	92.1
Lec_1_4	259	93	0.359	0.448	0.372	2.46	0.16	1.18	2.30	1.25	2.00	0.41	0.53	71.0	74.4
Lec_1_5	259	240	0.927	0.031	0.281	-1.42	0.25	1.00	0.10	1.17	0.60	0.25	0.26	91.6	92.1
Lec_1_6	259	234	0.903	0.139	0.422	-1.09	0.22	0.90	-0.60	0.62	-1.40	0.38	0.29	90.8	89.7
Lec_1_7	259	171	0.660	0.402	0.401	0.78	0.15	1.08	1.30	1.15	1.60	0.39	0.45	69.7	70.1
Lec_1_8	259	232	0.896	0.170	0.412	-0.99	0.22	0.92	-0.50	0.73	-1.00	0.36	0.29	89.1	88.9
Lec_1_9	259	187	0.722	0.432	0.483	0.41	0.15	0.95	-0.70	0.89	-0.90	0.47	0.43	74.8	73.1
Lec_1_10	259	210	0.811	0.309	0.475	-0.19	0.17	0.92	-0.90	0.75	-1.60	0.45	0.37	81.5	80.5
Lec_1_11	259	179	0.691	0.355	0.433	0.60	0.15	1.02	0.30	1.02	0.30	0.43	0.44	68.9	71.3
Lec_1_12	259	168	0.649	0.463	0.455	0.84	0.15	1.00	0.10	1.04	0.50	0.45	0.46	71.0	69.6
Lec_1_13	259	154	0.595	0.479	0.501	1.14	0.14	0.96	-0.70	0.92	-1.00	0.51	0.48	69.3	68.6
Lec_1_14	259	192	0.741	0.355	0.400	0.29	0.16	1.03	0.40	1.00	0.00	0.40	0.41	72.7	74.5
Lec_1_15	259	195	0.753	0.355	0.458	0.22	0.16	0.96	-0.50	0.84	-1.30	0.45	0.41	76.5	75.4
Lec_2_1	267	250	0.936	0.015	0.257	-2.59	0.26	1.01	0.10	0.83	-0.30	0.25	0.25	93.6	93.6
Lec_2_2	267	206	0.772	0.390	0.524	-0.88	0.16	0.88	-1.40	0.97	-0.10	0.49	0.41	80.8	79.7
Lec_2_3	267	177	0.663	0.390	0.444	-0.20	0.15	1.03	0.50	1.08	0.70	0.43	0.46	70.9	73.5
Lec_2_4	267	228	0.854	0.300	0.484	-1.55	0.19	0.88	-1.10	0.61	-1.70	0.45	0.35	84.9	85.7
Lec_2_5	267	117	0.438	0.599	0.503	1.01	0.14	0.96	-0.70	1.06	0.70	0.50	0.49	75.5	71.6
Lec_2_6	267	227	0.850	0.225	0.363	-1.51	0.19	1.02	0.20	0.89	-0.40	0.35	0.35	86.8	85.4
Lec_2_7	267	145	0.543	0.584	0.476	0.45	0.14	1.01	0.20	0.98	-0.20	0.48	0.48	70.2	71.0
Lec_2_8	267	50	0.187	0.240	0.261	2.59	0.18	1.08	0.80	1.69	2.80	0.31	0.43	84.5	83.0
Lec_2_9	267	211	0.793	0.316	0.369	-1.02	0.17	1.07	0.80	0.97	-0.10	0.36	0.40	80.0	80.8
Lec_2_10	267	178	0.667	0.509	0.508	-0.22	0.15	0.95	-0.70	0.90	-0.80	0.49	0.46	75.1	73.7
Lec_2_11	267	162	0.607	0.554	0.511	0.11	0.14	0.96	-0.70	0.92	-0.70	0.50	0.47	72.8	71.5

Lec_2_12	267	178	0.667	0.569	0.567	-0.22	0.15	0.88	-1.90	0.77	-2.00	0.55	0.46	76.6	73.7
Lec_2_13	267	44	0.165	0.180	0.199	2.79	0.18	1.26	2.20	1.56	2.10	0.21	0.41	83.8	85.1
Lec_2_14	267	183	0.685	0.524	0.513	-0.33	0.15	0.93	-1.00	0.93	-0.50	0.49	0.45	73.6	74.6
Lec_2_15	267	90	0.338	0.481	0.405	1.57	0.15	1.05	0.70	1.18	1.50	0.43	0.48	74.7	74.7
Exam_1	306	177	0.578	0.340	0.354	-0.45	0.13	1.12	2.40	1.23	2.60	0.35	0.45	63.9	69.3
Exam_2	306	144	0.471	0.353	0.385	0.10	0.13	1.10	2.00	1.09	1.20	0.39	0.46	64.5	68.7
Exam_3	306	166	0.542	0.418	0.396	-0.27	0.13	1.08	1.60	1.14	1.70	0.39	0.45	65.6	69.0
Exam_4	306	165	0.539	0.575	0.545	-0.25	0.13	0.90	-2.20	0.89	-1.50	0.53	0.45	73.9	68.9
Exam_5	306	95	0.310	0.497	0.471	0.97	0.14	0.98	-0.30	0.94	-0.50	0.47	0.45	75.3	74.9
Exam_6	306	73	0.239	0.340	0.355	1.43	0.15	1.06	0.80	1.50	3.00	0.36	0.44	78.9	80.0
Exam_7	306	238	0.780	0.367	0.498	-1.63	0.15	0.89	-1.30	0.73	-1.80	0.49	0.40	83.9	79.8
Exam_8	306	197	0.644	0.444	0.510	-0.80	0.13	0.93	-1.30	0.84	-1.60	0.50	0.44	72.6	71.9
Exam_9	306	198	0.647	0.549	0.556	-0.82	0.13	0.87	-2.50	0.76	-2.60	0.55	0.44	75.6	72.0
Exam_10	306	61	0.199	0.314	0.373	1.72	0.16	1.04	0.40	1.12	0.70	0.39	0.43	83.9	82.7
Red_Exam_1	258	244	0.942	0.201	0.284	-2.52	0.28	0.97	-0.10	1.25	0.70	0.23	0.21	94.6	94.5
Red_Exam_2	258	153	0.593	0.589	0.431	0.31	0.14	0.99	-0.10	0.98	-0.20	0.42	0.41	66.9	68.6
Red_Exam_3	258	239	0.923	0.185	0.235	-2.17	0.25	1.01	0.10	1.21	0.70	0.20	0.24	93.0	92.6
Red_Exam_4	258	153	0.591	0.448	0.293	0.31	0.14	1.14	2.50	1.18	2.20	0.29	0.41	61.5	68.6
Red_Exam_5	258	233	0.900	0.201	0.192	-1.84	0.22	1.09	0.60	1.62	2.00	0.15	0.27	90.3	90.3
Red_Exam_6	258	247	0.954	0.185	0.279	-2.79	0.32	0.95	-0.10	1.53	1.20	0.22	0.19	95.7	95.7
Red_Exam_7	258	242	0.934	0.247	0.344	-2.37	0.27	0.95	-0.20	0.64	-1.00	0.30	0.22	93.8	93.8
Red_Exam_8	258	203	0.784	0.556	0.452	-0.79	0.16	0.93	-0.80	0.82	-1.20	0.43	0.35	80.5	79.9
Red_Exam_9	258	215	0.830	0.602	0.503	-1.14	0.18	0.87	-1.20	0.62	-2.20	0.47	0.33	83.3	83.8
Red_Exam_10	258	238	0.919	0.309	0.367	-2.11	0.24	0.92	-0.40	0.78	-0.60	0.32	0.24	92.6	92.2
Red_Exam_11	258	161	0.622	0.664	0.463	0.16	0.14	0.96	-0.70	0.89	-1.30	0.45	0.41	66.1	69.6
Red_Exam_12	258	143	0.552	0.602	0.354	0.51	0.14	1.07	1.40	1.13	1.80	0.34	0.42	66.1	67.9
Red_Exam_13	258	132	0.514	0.560	0.317	0.71	0.14	1.11	2.20	1.19	2.50	0.31	0.42	63.8	67.7
Red_Exam_14	258	169	0.653	0.602	0.423	-0.01	0.14	1.00	0.00	0.93	-0.70	0.41	0.40	70.0	71.0
Red_Exam_15	258	74	0.287	-0.124	-0.059	1.88	0.15	1.46	5.70	2.06	6.90	-0.05	0.40	64.2	74.9
Red_Exam_16	258	106	0.409	0.757	0.489	1.21	0.14	0.90	-1.90	0.96	-0.40	0.49	0.42	74.7	69.5

Red_Exam_17	258	128	0.494	0.710	0.464	0.79	0.14	0.95	-0.90	0.93	-0.90	0.46	0.42	70.8	67.8
Red_Exam_18	258	81	0.314	0.326	0.231	1.72	0.15	1.16	2.40	1.33	2.80	0.24	0.40	70.4	73.4
Red_Exam_19	258	197	0.761	0.510	0.403	-0.64	0.16	0.97	-0.40	0.98	-0.10	0.38	0.36	79.4	78.1
Red_Exam_20	258	156	0.602	0.695	0.476	0.25	0.14	0.94	-1.10	0.95	-0.60	0.46	0.41	72.8	68.9
Red_Exam_21	258	139	0.537	0.618	0.403	0.58	0.14	1.03	0.60	1.05	0.70	0.39	0.42	66.9	67.8
Red_Exam_22	258	127	0.490	0.741	0.500	0.81	0.14	0.92	-1.50	0.88	-1.70	0.49	0.42	71.2	67.8
Red_Exam_23	258	187	0.725	0.713	0.489	-0.40	0.15	0.91	-1.30	0.82	-1.60	0.47	0.38	79.4	75.3
Red_Exam_24	258	115	0.446	0.543	0.364	1.04	0.14	1.05	1.00	1.11	1.40	0.37	0.42	64.2	68.7
Red_Exam_25	258	171	0.660	0.664	0.474	-0.05	0.14	0.94	-1.00	0.90	-1.10	0.46	0.40	73.2	71.4
Red_Exam_26	258	91	0.351	0.633	0.433	1.51	0.14	0.96	-0.60	0.95	-0.50	0.44	0.41	73.9	71.5
Red_Exam_27	258	109	0.421	0.757	0.501	1.15	0.14	0.91	-1.70	0.87	-1.70	0.50	0.42	72.8	69.2
Red_Exam_28	258	138	0.535	0.884	0.564	0.60	0.14	0.85	-3.10	0.82	-2.70	0.55	0.42	74.3	67.7
Red_Exam_29	258	50	0.193	0.263	0.245	2.49	0.17	1.07	0.70	1.29	1.60	0.27	0.36	80.9	81.6
Red_Exam_30	258	130	0.504	0.915	0.580	0.75	0.14	0.83	-3.50	0.82	-2.70	0.57	0.42	75.5	67.8

Table 58: Breakdown of Individual Items used in Foundations of Chemistry IA 2013 Multiple-Choice Assessments using CTT and Rasch Analysis

Foundations of Chemistry IA 2013															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	309	283	0.916	0.466	0.228	-1.34	0.22	1.07	0.50	1.31	1.00	0.22	0.28	91.5	91.3
Lec_1_2	309	209	0.676	0.932	0.521	0.62	0.14	0.92	-1.40	0.88	-1.10	0.51	0.45	76.1	72.4
Lec_1_3	309	284	0.919	0.453	0.186	-1.39	0.22	1.12	0.70	1.61	1.70	0.18	0.27	91.1	91.6
Lec_1_4	309	91	0.296	0.691	0.454	2.80	0.15	0.99	-0.20	1.10	0.70	0.51	0.52	79.5	78.1
Lec_1_5	309	276	0.893	0.595	0.394	-1.05	0.20	0.94	-0.40	0.79	-0.70	0.35	0.31	89.1	89.0
Lec_1_6	309	283	0.916	0.544	0.354	-1.34	0.22	0.96	-0.20	0.69	-1.00	0.32	0.28	91.5	91.3
Lec_1_7	309	215	0.696	0.803	0.409	0.50	0.14	1.06	1.00	1.11	1.10	0.40	0.44	72.0	73.0
Lec_1_8	309	271	0.877	0.621	0.379	-0.86	0.19	0.97	-0.20	0.85	-0.60	0.35	0.32	87.4	87.4
Lec_1_9	309	228	0.743	0.782	0.470	0.24	0.14	0.97	-0.40	0.85	-1.20	0.46	0.42	74.4	75.5
Lec_1_10	309	271	0.877	0.544	0.317	-0.86	0.19	1.02	0.20	1.06	0.30	0.30	0.32	87.4	87.4
Lec_1_11	309	240	0.777	0.880	0.567	-0.02	0.15	0.83	-2.30	0.71	-2.10	0.52	0.40	81.6	78.2
Lec_1_12	309	178	0.580	0.834	0.459	1.18	0.13	1.06	1.10	1.10	1.20	0.44	0.49	69.3	70.1
Lec_1_13	309	129	0.424	0.882	0.473	2.04	0.14	1.02	0.30	1.07	0.80	0.50	0.52	72.7	72.3
Lec_1_14	309	228	0.738	0.634	0.365	0.24	0.14	1.09	1.40	1.12	1.00	0.36	0.42	71.0	75.5
Lec_1_15	309	268	0.867	0.621	0.369	-0.76	0.18	0.99	0.00	0.82	-0.80	0.35	0.33	87.0	86.5
Lec_2_1	255	239	0.937	0.408	0.253	-2.89	0.27	1.01	0.10	1.15	0.50	0.25	0.26	94.1	93.8
Lec_2_2	255	191	0.749	0.800	0.520	-0.98	0.16	0.91	-1.10	0.76	-1.40	0.50	0.43	79.4	78.2
Lec_2_3	255	156	0.612	0.831	0.535	-0.16	0.15	0.93	-1.10	0.91	-0.80	0.52	0.47	75.5	72.0
Lec_2_4	255	194	0.761	0.769	0.532	-1.06	0.16	0.88	-1.50	0.75	-1.40	0.51	0.42	81.8	79.0
Lec_2_5	255	106	0.416	0.659	0.442	0.89	0.15	1.04	0.60	1.15	1.40	0.45	0.49	73.5	72.0
Lec_2_6	255	195	0.765	0.753	0.492	-1.09	0.17	0.93	-0.80	0.85	-0.80	0.47	0.42	79.1	79.3
Lec_2_7	255	132	0.518	0.722	0.469	0.34	0.14	1.03	0.50	1.03	0.30	0.47	0.48	70.8	71.0
Lec_2_8	255	37	0.145	0.329	0.300	2.74	0.20	1.02	0.20	1.42	1.50	0.34	0.40	87.4	86.9
Lec_2_9	255	206	0.808	0.486	0.289	-1.40	0.18	1.15	1.50	1.32	1.40	0.29	0.39	80.6	82.2
Lec_2_10	255	159	0.624	0.957	0.583	-0.23	0.15	0.86	-2.40	0.97	-0.20	0.55	0.47	78.3	72.2
Lec_2_11	255	152	0.598	0.724	0.434	-0.08	0.15	1.07	1.20	1.02	0.20	0.43	0.47	69.2	71.7

Lec_2_12	255	159	0.624	0.878	0.530	-0.23	0.15	0.93	-1.10	0.90	-0.80	0.51	0.47	75.9	72.2
Lec_2_13	255	41	0.161	0.031	0.030	2.59	0.19	1.41	3.20	2.78	4.90	0.06	0.41	84.2	85.4
Lec_2_14	255	158	0.620	0.957	0.600	-0.20	0.15	0.85	-2.60	0.73	-2.60	0.58	0.47	77.9	72.1
Lec_2_15	255	69	0.271	0.612	0.467	1.74	0.16	0.93	-0.90	1.05	0.40	0.49	0.46	81.4	78.0
Exam_1	365	179	0.490	0.318	0.410	-0.12	0.12	1.09	1.90	1.09	1.20	0.41	0.47	64.4	69.6
Exam_2	365	155	0.425	0.395	0.444	0.22	0.12	1.06	1.20	1.10	1.30	0.43	0.48	70.6	71.1
Exam_3	365	198	0.544	0.473	0.493	-0.39	0.12	0.97	-0.60	0.96	-0.60	0.49	0.47	70.3	69.6
Exam_4	365	219	0.600	0.460	0.530	-0.69	0.12	0.91	-1.80	0.87	-1.60	0.52	0.45	73.7	70.4
Exam_5	365	107	0.293	0.373	0.423	0.98	0.13	1.05	0.80	1.15	1.30	0.43	0.47	75.9	76.7
Exam_6	365	101	0.277	0.340	0.377	1.09	0.13	1.08	1.20	1.28	2.30	0.40	0.47	79.6	77.7
Exam_7	365	275	0.753	0.230	0.467	-1.59	0.13	0.93	-1.00	0.80	-1.50	0.46	0.40	79.0	77.4
Exam_8	365	229	0.627	0.362	0.470	-0.84	0.12	0.96	-0.80	0.96	-0.40	0.47	0.45	74.2	71.2
Exam_9	365	232	0.636	0.515	0.591	-0.89	0.12	0.83	-3.60	0.73	-3.20	0.57	0.44	79.0	71.5
Exam_10	365	50	0.137	0.231	0.319	2.24	0.17	1.11	1.00	1.40	1.70	0.33	0.43	86.3	88.1
Red_Exam_1	346	313	0.858	0.395	0.604	-2.17	0.23	1.03	0.20	2.00	2.50	0.46	0.51	93.4	93.4
Red_Exam_2	346	149	0.408	0.603	0.435	1.11	0.12	1.03	0.60	1.02	0.30	0.39	0.41	66.6	68.7
Red_Exam_3	346	325	0.890	0.384	0.702	-3.03	0.33	0.98	0.00	0.68	-0.70	0.59	0.57	97.0	97.0
Red_Exam_4	346	271	0.742	0.405	0.461	-0.86	0.15	1.16	1.90	1.36	1.90	0.34	0.44	78.2	81.2
Red_Exam_5	346	294	0.805	0.570	0.645	-1.44	0.17	0.91	-0.70	0.75	-1.10	0.52	0.47	87.8	87.8
Red_Exam_6	346	318	0.871	0.460	0.682	-2.46	0.25	0.97	-0.10	0.68	-0.90	0.56	0.53	94.9	94.9
Red_Exam_7	346	320	0.877	0.427	0.696	-2.59	0.27	0.95	-0.10	0.61	-1.10	0.57	0.54	95.5	95.5
Red_Exam_8	346	264	0.723	0.668	0.597	-0.71	0.14	0.94	-0.70	0.81	-1.30	0.49	0.44	77.9	79.3
Red_Exam_9	346	271	0.745	0.681	0.643	-0.86	0.15	0.88	-1.50	0.68	-2.00	0.53	0.44	82.4	81.2
Red_Exam_10	346	307	0.841	0.537	0.672	-1.89	0.20	0.94	-0.40	0.63	-1.40	0.54	0.49	91.6	91.6
Red_Exam_11	346	235	0.644	0.690	0.562	-0.18	0.13	0.95	-0.90	0.84	-1.40	0.47	0.43	71.6	72.7
Red_Exam_12	346	176	0.482	0.526	0.382	0.71	0.12	1.16	3.60	1.31	4.00	0.31	0.42	59.1	67.3
Red_Exam_13	346	213	0.585	0.703	0.506	0.17	0.12	1.00	0.10	1.01	0.10	0.42	0.43	71.3	69.4
Red_Exam_14	346	246	0.674	0.625	0.546	-0.37	0.13	0.99	-0.10	0.88	-0.90	0.45	0.43	74.3	75.0
Red_Exam_15	346	164	0.449	0.614	0.446	0.89	0.12	1.04	0.80	1.03	0.40	0.39	0.41	65.4	67.7
Red_Exam_16	347	120	0.329	0.493	0.408	1.56	0.13	1.04	0.70	1.04	0.50	0.37	0.39	69.3	72.3

Red_Exam_17	347	163	0.447	0.570	0.472	0.91	0.12	0.99	-0.30	0.98	-0.30	0.42	0.41	70.2	67.7
Red_Exam_18	347	78	0.214	0.263	0.261	2.31	0.14	1.15	1.80	1.38	2.90	0.25	0.36	78.3	80.0
Red_Exam_19	347	262	0.718	0.658	0.586	-0.65	0.14	0.94	-0.80	0.81	-1.30	0.48	0.44	78.6	78.7
Red_Exam_20	347	214	0.586	0.690	0.524	0.16	0.12	0.96	-0.80	0.93	-0.70	0.45	0.42	71.4	69.5
Red_Exam_21	347	189	0.518	0.548	0.432	0.53	0.12	1.08	1.80	1.16	2.00	0.36	0.42	68.5	67.4
Red_Exam_22	347	171	0.468	0.592	0.466	0.79	0.12	1.00	0.10	1.01	0.20	0.41	0.41	68.8	67.3
Red_Exam_23	347	228	0.625	0.734	0.562	-0.06	0.13	0.92	-1.50	0.84	-1.60	0.48	0.43	73.5	71.4
Red_Exam_24	347	173	0.475	0.549	0.421	0.77	0.12	1.07	1.60	1.04	0.60	0.38	0.41	61.6	67.3
Red_Exam_25	347	211	0.578	0.756	0.551	0.21	0.12	0.92	-1.80	0.84	-1.90	0.48	0.42	71.7	69.1
Red_Exam_26	347	122	0.334	0.537	0.402	1.53	0.13	1.06	1.00	1.09	1.10	0.36	0.40	70.5	72.1
Red_Exam_27	347	140	0.384	0.581	0.432	1.25	0.12	1.03	0.60	1.04	0.60	0.39	0.40	68.5	69.6
Red_Exam_28	347	177	0.485	0.855	0.560	0.71	0.12	0.87	-3.10	0.84	-2.40	0.49	0.42	74.1	67.3
Red_Exam_29	347	47	0.129	0.197	0.200	3.06	0.17	1.13	1.10	1.77	3.40	0.20	0.32	87.2	87.2
Red_Exam_30	347	185	0.507	0.800	0.564	0.59	0.12	0.86	-3.40	0.80	-2.80	0.50	0.42	73.2	67.3

Table 59: Breakdown of Individual Items used in Foundations of Chemistry IA 2014 Multiple-Choice Assessments using CTT and Rasch Analysis

Foundations of Chemistry IA 2014															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	252	231	0.917	-0.016	0.237	-1.45	0.25	1.12	0.70	1.70	1.70	0.22	0.33	90.7	91.5
Lec_1_2	252	192	0.762	0.302	0.425	0.07	0.17	1.06	0.70	1.04	0.30	0.42	0.45	77.1	78.3
Lec_1_3	252	225	0.893	0.159	0.510	-1.12	0.22	0.84	-1.10	0.63	-1.30	0.45	0.35	91.5	89.3
Lec_1_4	252	74	0.294	0.476	0.371	2.90	0.17	1.12	1.40	1.70	3.20	0.43	0.52	77.1	78.6
Lec_1_5	252	230	0.913	0.127	0.491	-1.39	0.24	0.84	-0.90	0.57	-1.30	0.43	0.33	91.9	91.1
Lec_1_6	252	225	0.893	0.143	0.424	-1.12	0.22	0.97	-0.10	0.74	-0.80	0.39	0.35	89.0	89.3
Lec_1_7	252	177	0.702	0.397	0.451	0.47	0.16	1.07	0.90	1.12	1.00	0.43	0.48	72.9	74.3
Lec_1_8	252	204	0.810	0.302	0.482	-0.29	0.18	0.97	-0.30	0.83	-0.90	0.46	0.42	82.2	82.0
Lec_1_9	252	178	0.706	0.444	0.547	0.44	0.16	0.92	-1.20	0.79	-1.70	0.54	0.48	73.7	74.5
Lec_1_10	252	216	0.857	0.190	0.421	-0.72	0.20	0.99	-0.10	1.03	0.20	0.40	0.39	85.2	86.1
Lec_1_11	252	199	0.790	0.270	0.475	-0.14	0.17	0.98	-0.20	0.98	0.00	0.45	0.44	81.8	80.5
Lec_1_12	252	166	0.659	0.540	0.555	0.73	0.15	0.93	-1.10	0.90	-0.90	0.54	0.50	75.8	72.7
Lec_1_13	252	120	0.476	0.651	0.542	1.77	0.15	0.95	-0.80	0.91	-0.90	0.56	0.53	72.5	71.4
Lec_1_14	252	169	0.671	0.365	0.403	0.66	0.15	1.12	1.70	1.15	1.30	0.42	0.49	68.6	73.2
Lec_1_15	252	218	0.865	0.159	0.364	-0.80	0.20	1.06	0.50	1.34	1.30	0.34	0.38	86.0	86.8
Lec_2_1	223	212	0.951	0.287	0.358	-2.88	0.32	0.87	-0.40	0.41	-1.40	0.33	0.21	95.1	95.1
Lec_2_2	223	176	0.789	0.717	0.566	-1.03	0.18	0.80	-2.20	0.61	-2.10	0.53	0.38	83.9	80.5
Lec_2_3	223	148	0.664	0.610	0.399	-0.25	0.16	1.06	0.90	1.02	0.20	0.40	0.43	72.2	73.1
Lec_2_4	223	189	0.848	0.556	0.441	-1.50	0.20	0.92	-0.60	0.86	-0.40	0.41	0.33	83.4	85.0
Lec_2_5	223	124	0.556	0.843	0.559	0.32	0.15	0.88	-2.10	0.84	-1.60	0.55	0.46	76.2	69.8
Lec_2_6	223	181	0.812	0.484	0.409	-1.20	0.19	0.97	-0.20	0.80	-0.90	0.40	0.36	83.4	82.2
Lec_2_7	223	123	0.552	0.664	0.416	0.35	0.15	1.05	0.80	1.12	1.20	0.42	0.46	68.6	69.8
Lec_2_8	223	48	0.215	0.287	0.260	2.27	0.18	1.10	1.00	1.75	3.30	0.30	0.43	82.1	81.5
Lec_2_9	223	190	0.852	0.413	0.280	-1.54	0.20	1.07	0.60	1.17	0.70	0.27	0.33	86.5	85.4
Lec_2_10	223	149	0.668	0.825	0.551	-0.28	0.16	0.86	-2.10	0.83	-1.30	0.53	0.43	80.7	73.3
Lec_2_11	223	120	0.538	0.628	0.438	0.42	0.15	1.03	0.60	1.05	0.60	0.43	0.46	71.7	69.6

Lec_2_12	223	130	0.583	0.646	0.419	0.18	0.15	1.05	0.90	1.01	0.10	0.42	0.45	66.4	70.1
Lec_2_13	223	30	0.135	0.126	0.095	2.97	0.22	1.29	2.00	2.31	3.50	0.11	0.38	85.7	87.5
Lec_2_14	223	125	0.561	0.807	0.542	0.30	0.15	0.91	-1.60	0.84	-1.60	0.53	0.46	74.0	69.9
Lec_2_15	223	61	0.274	0.502	0.372	1.87	0.17	1.01	0.20	1.33	2.00	0.41	0.45	76.7	77.6
Exam_1	327	177	0.541	0.587	0.431	-0.54	0.13	1.06	1.30	1.05	0.60	0.43	0.47	68.4	69.7
Exam_2	327	149	0.456	0.563	0.412	-0.09	0.13	1.10	2.00	1.11	1.30	0.42	0.48	65.6	70.3
Exam_3	327	184	0.563	0.709	0.496	-0.65	0.13	0.98	-0.40	0.91	-0.90	0.49	0.46	68.1	69.8
Exam_4	327	168	0.514	0.807	0.554	-0.40	0.13	0.90	-2.00	0.86	-1.60	0.54	0.47	76.2	70.0
Exam_5	327	103	0.315	0.599	0.502	0.70	0.14	0.97	-0.40	0.96	-0.40	0.50	0.49	77.4	76.4
Exam_6	327	72	0.220	0.343	0.342	1.35	0.15	1.14	1.50	1.52	2.90	0.36	0.47	80.8	82.2
Exam_7	327	234	0.716	0.673	0.492	-1.50	0.14	0.92	-1.30	1.03	0.30	0.47	0.42	76.5	75.6
Exam_8	327	177	0.541	0.648	0.488	-0.54	0.13	0.98	-0.30	0.92	-0.90	0.49	0.47	69.7	69.7
Exam_9	327	209	0.641	0.847	0.591	-1.06	0.13	0.82	-3.60	0.69	-3.00	0.58	0.44	77.7	71.6
Exam_10	327	29	0.089	0.110	0.202	2.74	0.22	1.12	0.80	2.61	3.50	0.23	0.39	91.6	91.8
Red_Exam_1	300	274	0.838	0.502	0.589	-2.01	0.22	1.00	0.00	1.30	1.00	0.21	0.25	91.6	91.5
Red_Exam_2	300	137	0.419	0.771	0.551	0.98	0.13	0.89	-2.20	0.84	-2.00	0.53	0.44	73.2	69.1
Red_Exam_3	300	282	0.862	0.538	0.668	-2.45	0.25	0.90	-0.40	0.83	-0.30	0.29	0.22	94.3	94.1
Red_Exam_4	300	209	0.639	0.477	0.430	-0.26	0.14	1.19	3.00	1.36	2.50	0.23	0.39	67.6	73.4
Red_Exam_5	300	269	0.823	0.563	0.612	-1.80	0.20	0.98	-0.10	0.98	0.00	0.27	0.27	90.0	89.9
Red_Exam_6	300	281	0.859	0.502	0.624	-2.38	0.25	0.99	0.00	1.04	0.20	0.21	0.22	94.0	93.8
Red_Exam_7	300	288	0.881	0.489	0.683	-2.91	0.30	0.91	-0.30	0.80	-0.30	0.26	0.19	96.3	96.1
Red_Exam_8	300	229	0.702	0.736	0.584	-0.67	0.15	0.95	-0.60	0.87	-0.70	0.40	0.36	79.9	77.8
Red_Exam_9	300	204	0.624	0.820	0.608	-0.17	0.14	0.89	-2.00	0.80	-1.70	0.48	0.39	75.9	72.5
Red_Exam_10	300	267	0.817	0.599	0.605	-1.72	0.19	1.00	0.00	0.91	-0.20	0.28	0.27	89.3	89.2
Red_Exam_11	300	208	0.636	0.807	0.594	-0.24	0.14	0.90	-1.70	0.83	-1.40	0.47	0.39	77.3	73.2
Red_Exam_12	300	155	0.474	0.514	0.397	0.67	0.13	1.17	3.40	1.19	2.10	0.31	0.44	61.2	68.4
Red_Exam_13	300	198	0.606	0.661	0.491	-0.06	0.13	1.07	1.30	1.08	0.70	0.35	0.40	67.2	71.6
Red_Exam_14	299	229	0.700	0.661	0.565	-0.68	0.15	0.99	-0.10	0.93	-0.30	0.37	0.36	77.9	78.0
Red_Exam_15	299	160	0.489	0.746	0.519	0.59	0.13	0.97	-0.50	0.96	-0.50	0.45	0.43	73.2	68.3
Red_Exam_16	299	112	0.343	0.624	0.475	1.41	0.13	0.95	-0.90	0.96	-0.40	0.49	0.45	73.5	71.8

Red_Exam_17	299	162	0.495	0.746	0.510	0.56	0.13	1.00	0.00	1.14	1.50	0.42	0.43	68.1	68.4
Red_Exam_18	299	65	0.199	0.281	0.252	2.39	0.16	1.29	3.00	1.59	3.20	0.21	0.44	77.2	81.8
Red_Exam_19	299	250	0.765	0.709	0.634	-1.19	0.17	0.92	-0.80	0.72	-1.30	0.40	0.32	84.9	84.1
Red_Exam_20	299	175	0.535	0.795	0.562	0.34	0.13	0.94	-1.20	0.87	-1.40	0.48	0.42	72.1	69.2
Red_Exam_21	299	170	0.520	0.758	0.517	0.43	0.13	0.99	-0.30	1.07	0.80	0.43	0.43	69.5	68.7
Red_Exam_22	299	153	0.468	0.893	0.591	0.71	0.13	0.84	-3.60	0.76	-3.10	0.57	0.44	75.5	68.4
Red_Exam_23	299	200	0.612	0.820	0.583	-0.10	0.14	0.92	-1.50	0.85	-1.20	0.46	0.40	74.5	72.0
Red_Exam_24	299	143	0.437	0.612	0.450	0.88	0.13	1.05	1.10	1.04	0.50	0.40	0.44	67.4	68.8
Red_Exam_25	299	195	0.596	0.771	0.528	-0.01	0.13	1.01	0.20	0.94	-0.50	0.40	0.40	70.8	71.3
Red_Exam_26	299	73	0.223	0.440	0.366	2.20	0.15	1.06	0.70	1.13	0.90	0.39	0.44	78.5	79.9
Red_Exam_27	299	108	0.330	0.575	0.426	1.49	0.14	1.03	0.50	1.11	1.10	0.42	0.45	72.8	72.5
Red_Exam_28	299	142	0.434	0.795	0.538	0.90	0.13	0.90	-2.10	0.87	-1.60	0.52	0.44	74.5	68.8
Red_Exam_29	299	64	0.196	0.171	0.176	2.41	0.16	1.30	3.00	3.07	8.50	0.14	0.43	76.8	82.0
Red_Exam_30	299	155	0.474	0.942	0.586	0.68	0.13	0.84	-3.50	0.79	-2.70	0.56	0.44	74.2	68.4

Table 60: Breakdown of Individual Items used in Foundations of Chemistry IA 2015 Multiple-Choice Assessments using CTT and Rasch Analysis

Foundations of Chemistry IA 2015															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	294	266	0.905	0.054	0.343	-1.30	0.21	1.00	0.10	0.86	-0.40	0.32	0.32	91.5	89.9
Lec_1_2	294	199	0.677	0.395	0.428	0.59	0.14	1.10	1.50	1.03	0.40	0.44	0.49	68.6	73.3
Lec_1_3	294	269	0.915	-0.014	0.291	-1.44	0.22	1.04	0.30	1.28	0.90	0.27	0.31	91.1	90.9
Lec_1_4	294	94	0.321	0.464	0.400	2.73	0.16	1.14	1.80	1.53	3.00	0.46	0.56	74.9	78.2
Lec_1_5	294	253	0.861	0.109	0.363	-0.79	0.19	1.04	0.40	1.15	0.70	0.34	0.37	86.7	85.9
Lec_1_6	294	268	0.912	0.095	0.403	-1.39	0.22	0.93	-0.40	0.87	-0.30	0.35	0.31	90.8	90.6
Lec_1_7	294	192	0.653	0.408	0.428	0.73	0.14	1.12	1.90	1.14	1.40	0.43	0.50	68.3	72.5
Lec_1_8	294	245	0.833	0.272	0.466	-0.53	0.17	0.96	-0.40	0.81	-0.90	0.43	0.40	83.8	83.4
Lec_1_9	294	215	0.731	0.517	0.560	0.24	0.15	0.89	-1.50	0.79	-1.70	0.53	0.47	78.6	75.8
Lec_1_10	294	257	0.874	0.163	0.476	-0.93	0.19	0.90	-0.80	0.68	-1.30	0.43	0.36	88.2	87.1
Lec_1_11	294	234	0.796	0.272	0.476	-0.22	0.16	0.97	-0.30	0.92	-0.40	0.45	0.43	80.8	80.0
Lec_1_12	294	196	0.667	0.544	0.557	0.65	0.14	0.93	-1.10	0.92	-0.80	0.53	0.50	76.4	73.0
Lec_1_13	294	121	0.412	0.653	0.544	2.13	0.14	1.00	0.00	0.94	-0.50	0.57	0.56	73.8	73.8
Lec_1_14	294	202	0.687	0.463	0.534	0.53	0.15	0.95	-0.80	0.93	-0.60	0.52	0.49	74.2	73.6
Lec_1_15	294	259	0.881	0.122	0.381	-1.00	0.20	1.01	0.20	0.83	-0.60	0.36	0.35	87.5	87.7
Lec_2_1	236	221	0.936	0.203	0.436	-2.73	0.29	0.81	-0.90	0.44	-1.40	0.42	0.30	94.4	93.8
Lec_2_2	236	191	0.809	0.458	0.525	-1.23	0.19	0.86	-1.30	0.76	-1.10	0.50	0.42	85.9	83.1
Lec_2_3	236	143	0.606	0.407	0.388	0.09	0.15	1.14	2.10	1.21	1.70	0.38	0.48	69.7	72.8
Lec_2_4	236	191	0.809	0.475	0.554	-1.23	0.19	0.84	-1.50	0.64	-1.70	0.53	0.42	84.2	83.1
Lec_2_5	236	101	0.428	0.695	0.641	1.04	0.15	0.76	-4.30	0.73	-2.50	0.64	0.49	82.9	71.3
Lec_2_6	236	173	0.733	0.525	0.547	-0.67	0.17	0.90	-1.20	0.74	-1.60	0.53	0.45	79.9	77.9
Lec_2_7	236	120	0.508	0.458	0.381	0.61	0.15	1.13	2.10	1.41	3.40	0.39	0.49	67.9	70.8
Lec_2_8	236	54	0.229	0.339	0.274	2.25	0.18	1.11	1.20	1.86	3.20	0.31	0.44	81.2	80.8
Lec_2_9	236	193	0.818	0.237	0.298	-1.30	0.19	1.16	1.40	1.53	2.00	0.29	0.41	81.6	83.7
Lec_2_10	236	149	0.631	0.678	0.638	-0.05	0.16	0.80	-3.20	0.69	-2.80	0.62	0.48	81.6	73.6
Lec_2_11	236	149	0.631	0.407	0.401	-0.05	0.16	1.12	1.70	1.14	1.10	0.40	0.48	68.8	73.6

Lec_2_12	236	135	0.574	0.579	0.527	0.27	0.15	0.95	-0.70	0.93	-0.60	0.52	0.49	75.6	72.0
Lec_2_13	236	40	0.169	0.153	0.204	2.73	0.20	1.25	2.00	2.11	3.20	0.22	0.41	82.9	85.0
Lec_2_14	236	196	0.834	0.443	0.526	-1.41	0.19	0.86	-1.20	0.67	-1.40	0.50	0.41	88.0	84.6
Lec_2_15	236	75	0.318	0.492	0.407	1.67	0.16	1.02	0.40	1.41	2.30	0.43	0.47	76.5	75.5
Exam_1	367	204	0.556	0.349	0.397	-0.39	0.12	1.10	2.10	1.11	1.40	0.40	0.47	64.9	70.1
Exam_2	367	168	0.458	0.578	0.499	0.12	0.12	0.97	-0.50	0.98	-0.30	0.49	0.48	71.3	70.1
Exam_3	367	203	0.553	0.381	0.409	-0.38	0.12	1.08	1.80	1.12	1.60	0.40	0.47	68.0	70.1
Exam_4	367	207	0.564	0.567	0.550	-0.43	0.12	0.90	-2.20	0.89	-1.40	0.53	0.46	76.3	70.1
Exam_5	367	100	0.272	0.414	0.408	1.18	0.13	1.05	0.70	1.12	1.00	0.42	0.47	78.0	77.7
Exam_6	366	102	0.279	0.426	0.423	1.13	0.13	1.01	0.20	1.13	1.10	0.44	0.47	78.8	77.2
Exam_7	367	289	0.787	0.349	0.504	-1.76	0.14	0.86	-2.00	0.72	-1.90	0.50	0.39	81.9	80.3
Exam_8	367	238	0.649	0.403	0.441	-0.89	0.12	1.01	0.20	0.99	0.00	0.44	0.45	73.0	72.3
Exam_9	367	238	0.649	0.589	0.588	-0.89	0.12	0.83	-3.50	0.74	-3.00	0.57	0.45	79.1	72.3
Exam_10	367	50	0.136	0.196	0.291	2.31	0.17	1.17	1.40	1.39	1.60	0.30	0.42	85.5	88.4
Red_Exam_1	331	300	0.820	0.536	0.637	-1.98	0.20	1.01	0.10	1.04	0.30	0.25	0.27	90.9	90.8
Red_Exam_2	331	145	0.396	0.699	0.492	1.02	0.12	0.98	-0.50	1.11	1.40	0.46	0.44	73.3	69.7
Red_Exam_3	331	309	0.844	0.503	0.660	-2.39	0.23	0.95	-0.20	1.62	1.70	0.23	0.24	93.9	93.4
Red_Exam_4	331	250	0.683	0.579	0.538	-0.66	0.14	1.06	0.90	1.25	1.70	0.31	0.37	77.6	77.6
Red_Exam_5	331	299	0.817	0.525	0.636	-1.94	0.20	1.00	0.00	1.03	0.20	0.26	0.28	90.6	90.6
Red_Exam_6	331	302	0.825	0.536	0.654	-2.06	0.20	0.96	-0.20	1.02	0.20	0.28	0.27	92.1	91.4
Red_Exam_7	331	315	0.861	0.525	0.693	-2.76	0.27	0.98	0.00	0.82	-0.40	0.24	0.21	95.5	95.2
Red_Exam_8	331	258	0.705	0.689	0.614	-0.82	0.14	0.95	-0.60	0.87	-0.80	0.41	0.36	81.5	79.4
Red_Exam_9	331	221	0.604	0.929	0.657	-0.15	0.13	0.83	-3.40	0.72	-3.00	0.55	0.41	77.3	72.2
Red_Exam_10	331	282	0.770	0.656	0.634	-1.39	0.17	0.95	-0.50	0.92	-0.30	0.35	0.32	86.4	85.6
Red_Exam_11	331	224	0.612	0.885	0.644	-0.20	0.13	0.84	-3.00	0.78	-2.20	0.53	0.40	80.0	72.7
Red_Exam_12	331	193	0.527	0.645	0.464	0.30	0.12	1.11	2.30	1.21	2.40	0.32	0.43	65.2	68.9
Red_Exam_13	331	190	0.519	0.503	0.396	0.34	0.12	1.23	4.50	1.40	4.40	0.23	0.43	59.4	68.7
Red_Exam_14	331	236	0.645	0.732	0.591	-0.40	0.13	0.96	-0.60	0.90	-0.80	0.42	0.39	75.2	74.8
Red_Exam_15	331	180	0.492	0.656	0.463	0.49	0.12	1.10	2.10	1.12	1.60	0.35	0.43	63.9	68.3
Red_Exam_16	328	125	0.342	0.689	0.489	1.33	0.13	0.94	-1.10	0.94	-0.70	0.48	0.44	74.9	71.4

Red_Exam_17	328	176	0.481	0.787	0.551	0.55	0.12	0.93	-1.50	0.95	-0.70	0.48	0.43	70.9	68.0
Red_Exam_18	328	69	0.189	0.306	0.280	2.37	0.15	1.15	1.70	1.46	2.80	0.26	0.41	80.1	81.2
Red_Exam_19	328	259	0.708	0.765	0.623	-0.87	0.15	0.97	-0.40	0.85	-1.00	0.39	0.35	80.1	80.0
Red_Exam_20	328	177	0.484	0.776	0.530	0.53	0.12	0.98	-0.50	0.95	-0.70	0.45	0.43	69.4	68.1
Red_Exam_21	328	175	0.478	0.645	0.457	0.56	0.12	1.11	2.40	1.14	1.80	0.34	0.43	63.9	68.0
Red_Exam_22	328	173	0.473	0.842	0.571	0.59	0.12	0.89	-2.40	0.88	-1.60	0.52	0.43	72.5	68.1
Red_Exam_23	328	215	0.587	0.831	0.617	-0.06	0.13	0.89	-2.10	0.81	-2.10	0.50	0.40	72.5	71.4
Red_Exam_24	328	165	0.451	0.557	0.441	0.71	0.12	1.10	2.20	1.13	1.80	0.35	0.44	65.1	68.4
Red_Exam_25	328	225	0.615	0.863	0.648	-0.23	0.13	0.85	-2.80	0.78	-2.20	0.52	0.39	77.7	73.0
Red_Exam_26	328	87	0.238	0.503	0.411	2.00	0.14	0.98	-0.20	0.97	-0.20	0.44	0.43	77.7	77.6
Red_Exam_27	328	116	0.317	0.590	0.429	1.48	0.13	1.03	0.50	1.05	0.60	0.41	0.44	71.9	72.7
Red_Exam_28	328	161	0.440	0.754	0.529	0.78	0.12	0.95	-1.10	0.94	-0.80	0.48	0.44	70.6	68.5
Red_Exam_29	328	58	0.158	0.120	0.175	2.64	0.16	1.26	2.60	1.96	4.40	0.13	0.40	82.6	83.9
Red_Exam_30	328	200	0.546	0.940	0.632	0.18	0.13	0.84	-3.50	0.74	-3.30	0.56	0.42	73.4	69.4

Table 61: Breakdown of Individual Items used in Foundations of Chemistry IB 2012 Multiple-Choice Assessments using CTT and Rasch Analysis

Foundations of Chemistry IB 2012															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	236	208	0.882	0.101	0.299	-1.48	0.22	1.11	0.80	1.27	0.90	0.28	0.35	87.0	88.0
Lec_1_2	236	52	0.218	0.286	0.348	3.05	0.20	1.31	2.60	1.57	1.90	0.39	0.55	77.6	84.2
Lec_1_3	236	212	0.899	-0.050	0.199	-1.69	0.23	1.14	0.90	2.14	2.40	0.19	0.33	88.8	89.6
Lec_1_4	236	178	0.756	0.420	0.563	-0.38	0.17	0.89	-1.40	0.68	-1.90	0.53	0.45	81.2	78.0
Lec_1_5	236	182	0.773	0.403	0.518	-0.50	0.18	0.92	-0.90	0.87	-0.60	0.48	0.44	81.2	79.1
Lec_1_6	236	205	0.866	0.168	0.453	-1.34	0.21	0.91	-0.70	0.70	-1.00	0.42	0.36	87.4	86.8
Lec_1_7	236	126	0.529	0.790	0.608	0.96	0.16	0.87	-2.00	0.81	-1.80	0.61	0.54	74.0	72.0
Lec_1_8	236	182	0.773	0.387	0.444	-0.50	0.18	1.03	0.40	1.03	0.20	0.42	0.44	79.4	79.1
Lec_1_9	236	102	0.437	0.672	0.515	1.55	0.16	1.03	0.40	1.01	0.10	0.54	0.56	71.3	73.5
Lec_1_10	236	166	0.705	0.489	0.488	-0.04	0.16	1.02	0.30	0.88	-0.70	0.48	0.48	74.4	75.3
Lec_1_11	236	146	0.620	0.658	0.621	0.48	0.16	0.85	-2.30	0.74	-2.30	0.60	0.51	77.1	72.7
Lec_1_12	236	165	0.702	0.605	0.595	-0.01	0.16	0.86	-1.90	0.75	-1.80	0.56	0.48	79.4	75.1
Lec_1_13	236	200	0.849	0.319	0.509	-1.13	0.20	0.89	-1.00	0.67	-1.20	0.46	0.38	86.1	84.9
Lec_1_14	236	101	0.436	0.542	0.424	1.58	0.16	1.14	1.80	1.29	2.20	0.47	0.56	71.7	73.6
Lec_1_15	236	183	0.777	0.319	0.390	-0.53	0.18	1.10	1.10	1.14	0.70	0.38	0.44	77.1	79.4
Lec_2_1	189	120	0.635	0.508	0.460	-0.66	0.17	1.04	0.60	0.95	-0.30	0.45	0.46	68.8	72.8
Lec_2_2	189	121	0.640	0.614	0.509	-0.69	0.17	0.95	-0.70	1.02	0.20	0.49	0.46	76.7	73.0
Lec_2_3	189	24	0.127	-0.106	-0.079	2.54	0.24	1.49	2.90	4.34	5.40	-0.08	0.37	85.7	87.8
Lec_2_4	**Unused Due to Error**														
Lec_2_5	189	155	0.820	0.360	0.497	-1.88	0.21	0.88	-1.00	0.73	-1.00	0.48	0.39	85.2	83.7
Lec_2_6	189	41	0.217	0.508	0.431	1.74	0.20	0.97	-0.20	1.10	0.50	0.45	0.44	82.5	81.6
Lec_2_7	189	89	0.471	0.487	0.380	0.21	0.17	1.14	2.00	1.27	2.20	0.38	0.49	64.0	71.0
Lec_2_8	189	110	0.582	0.720	0.589	-0.38	0.17	0.87	-2.00	0.75	-2.10	0.57	0.48	75.7	71.2
Lec_2_9	189	33	0.175	0.317	0.372	2.08	0.21	0.95	-0.30	1.33	1.20	0.40	0.41	88.9	84.4
Lec_2_10	189	101	0.534	0.804	0.622	-0.12	0.17	0.82	-2.80	0.73	-2.50	0.61	0.48	78.3	70.4
Lec_2_11	189	103	0.545	0.402	0.372	-0.18	0.17	1.16	2.30	1.24	1.90	0.37	0.48	64.6	70.6

Lec_2_12	189	79	0.418	0.698	0.572	0.50	0.17	0.89	-1.60	0.80	-1.70	0.57	0.49	75.1	72.3
Lec_2_13	189	128	0.677	0.550	0.508	-0.91	0.18	0.95	-0.60	0.93	-0.40	0.49	0.45	75.1	74.5
Lec_2_14	189	135	0.714	0.571	0.590	-1.13	0.18	0.82	-2.20	0.71	-1.70	0.57	0.44	81.0	76.5
Lec_2_15	189	135	0.714	0.508	0.463	-1.13	0.18	1.02	0.30	0.90	-0.50	0.44	0.44	74.6	76.5
Exam_1	266	230	0.858	0.194	0.402	-1.40	0.19	0.98	-0.10	0.93	-0.20	0.35	0.33	86.4	86.4
Exam_2	266	237	0.884	0.104	0.365	-1.68	0.21	0.98	-0.10	1.05	0.30	0.31	0.31	89.1	89.0
Exam_3	266	96	0.358	0.493	0.346	1.66	0.15	1.18	2.70	1.28	2.40	0.35	0.49	71.2	73.4
Exam_4	266	58	0.216	0.448	0.444	2.61	0.17	1.00	0.10	0.96	-0.10	0.47	0.47	82.5	82.5
Exam_5	266	208	0.776	0.239	0.365	-0.71	0.16	1.07	0.90	1.19	1.20	0.33	0.39	80.5	79.1
Exam_6	266	122	0.455	0.567	0.443	1.12	0.14	1.03	0.60	1.15	1.70	0.45	0.48	68.9	69.4
Exam_7	266	137	0.511	0.657	0.525	0.83	0.14	0.94	-1.00	0.89	-1.40	0.52	0.48	70.8	69.4
Exam_8	266	225	0.840	0.418	0.510	-1.22	0.18	0.89	-1.00	0.59	-2.10	0.46	0.35	82.5	84.7
Exam_9	266	240	0.899	0.090	0.358	-1.82	0.22	0.98	-0.10	0.78	-0.70	0.32	0.30	91.1	90.1
Exam_10	266	148	0.552	0.701	0.557	0.61	0.14	0.90	-1.90	0.84	-2.00	0.55	0.47	72.8	69.0
Red_Exam_1	249	240	0.899	0.390	0.559	-3.10	0.35	0.99	0.10	1.03	0.20	0.16	0.18	96.3	96.3
Red_Exam_2	249	226	0.846	0.449	0.539	-2.02	0.23	1.01	0.10	1.10	0.40	0.25	0.27	90.9	90.6
Red_Exam_3	249	201	0.753	0.644	0.558	-1.03	0.18	0.96	-0.40	0.97	-0.10	0.37	0.36	83.9	81.3
Red_Exam_4	249	224	0.839	0.554	0.597	-1.91	0.22	0.90	-0.60	0.78	-0.60	0.34	0.28	90.1	89.8
Red_Exam_5	249	164	0.614	0.764	0.548	-0.08	0.15	0.97	-0.40	0.97	-0.20	0.45	0.43	74.8	72.5
Red_Exam_6	249	177	0.663	0.764	0.551	-0.38	0.16	0.99	-0.10	0.88	-0.80	0.43	0.41	71.5	75.0
Red_Exam_7	249	186	0.697	0.584	0.474	-0.61	0.16	1.12	1.50	1.16	1.00	0.31	0.39	73.1	77.0
Red_Exam_8	249	216	0.809	0.599	0.567	-1.56	0.20	0.96	-0.20	0.85	-0.50	0.34	0.31	87.2	86.6
Red_Exam_9	249	68	0.255	0.479	0.375	2.08	0.17	1.11	1.30	1.46	2.70	0.41	0.50	77.7	79.0
Red_Exam_10	249	173	0.648	0.764	0.561	-0.29	0.15	0.97	-0.40	0.87	-0.90	0.45	0.42	74.8	74.1
Red_Exam_11	249	178	0.667	0.839	0.596	-0.41	0.16	0.89	-1.50	1.01	0.10	0.47	0.41	80.2	75.2
Red_Exam_12	249	207	0.775	0.749	0.649	-1.23	0.18	0.82	-1.80	0.63	-1.80	0.46	0.34	86.0	83.2
Red_Exam_13	249	142	0.532	0.839	0.544	0.40	0.15	0.96	-0.70	0.90	-1.00	0.49	0.46	69.8	69.7
Red_Exam_14	249	182	0.682	0.584	0.445	-0.51	0.16	1.15	1.90	1.46	2.60	0.28	0.40	72.7	76.1
Red_Exam_15	249	188	0.704	0.809	0.624	-0.66	0.16	0.86	-1.80	0.72	-1.70	0.49	0.39	82.2	77.5
Red_Exam_16	250	154	0.577	0.809	0.553	0.16	0.15	0.97	-0.40	0.99	0.00	0.46	0.45	73.7	71.0

Red_Exam_17	250	143	0.536	0.839	0.561	0.40	0.15	0.95	-0.90	0.91	-0.90	0.50	0.46	70.0	69.8
Red_Exam_18	250	140	0.524	0.734	0.533	0.46	0.15	0.99	-0.10	0.95	-0.40	0.47	0.46	72.0	69.6
Red_Exam_19	250	139	0.521	0.824	0.573	0.48	0.15	0.91	-1.50	0.88	-1.30	0.52	0.46	75.7	69.5
Red_Exam_20	250	188	0.709	0.755	0.593	-0.64	0.16	0.93	-0.80	0.88	-0.70	0.44	0.39	78.6	77.4
Red_Exam_21	250	57	0.213	0.390	0.334	2.41	0.18	1.18	1.70	1.39	1.90	0.38	0.50	80.2	81.8
Red_Exam_22	250	108	0.404	0.494	0.414	1.14	0.15	1.16	2.50	1.17	1.70	0.39	0.49	63.0	70.6
Red_Exam_23	250	134	0.502	0.734	0.500	0.59	0.14	1.04	0.60	1.10	1.10	0.44	0.47	67.5	69.3
Red_Exam_24	250	55	0.206	0.494	0.415	2.48	0.18	0.95	-0.50	1.27	1.30	0.50	0.50	83.5	82.5
Red_Exam_25	250	105	0.393	0.869	0.592	1.20	0.15	0.84	-2.80	0.82	-1.90	0.59	0.49	76.5	71.2
Red_Exam_26	250	77	0.288	0.554	0.431	1.85	0.16	1.07	0.90	1.07	0.60	0.46	0.50	73.7	77.0
Red_Exam_27	250	150	0.562	0.719	0.516	0.25	0.15	1.03	0.60	1.04	0.50	0.43	0.45	70.0	70.5
Red_Exam_28	250	162	0.607	0.824	0.564	-0.01	0.15	0.96	-0.70	0.99	0.00	0.46	0.44	76.1	72.1
Red_Exam_29	250	162	0.607	0.614	0.480	-0.01	0.15	1.10	1.60	1.09	0.70	0.37	0.44	66.3	72.1
Red_Exam_30	250	135	0.506	0.764	0.514	0.56	0.14	1.01	0.30	1.08	0.80	0.45	0.47	67.9	69.3

Table 62: Breakdown of Individual Items used in Foundations of Chemistry IB 2013 Multiple-Choice Assessments using CTT and Rasch Analysis

Foundations of Chemistry IB 2013															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	249	220	0.884	0.241	0.342	-1.72	0.21	1.02	0.20	0.85	-0.40	0.32	0.32	87.7	88.2
Lec_1_2	249	55	0.221	0.418	0.388	2.55	0.18	1.09	0.90	1.32	1.50	0.41	0.48	80.7	82.0
Lec_1_3	249	225	0.904	0.161	0.287	-1.96	0.23	1.01	0.10	0.96	0.00	0.28	0.30	90.2	90.1
Lec_1_4	249	185	0.743	0.466	0.395	-0.55	0.16	1.07	0.90	1.03	0.20	0.39	0.43	76.2	77.5
Lec_1_5	249	202	0.811	0.450	0.487	-1.04	0.18	0.89	-1.10	0.81	-0.90	0.46	0.39	86.5	82.1
Lec_1_6	249	223	0.896	0.257	0.437	-1.86	0.22	0.84	-1.10	0.69	-0.90	0.41	0.31	89.3	89.3
Lec_1_7	249	106	0.426	0.819	0.596	1.24	0.15	0.85	-2.40	0.82	-1.90	0.60	0.51	79.5	72.0
Lec_1_8	249	199	0.799	0.257	0.337	-0.95	0.18	1.09	0.90	1.21	1.00	0.33	0.40	81.1	81.2
Lec_1_9	249	99	0.398	0.610	0.470	1.40	0.15	1.02	0.40	1.14	1.30	0.49	0.51	71.3	72.7
Lec_1_10	249	132	0.530	0.675	0.517	0.68	0.15	0.98	-0.40	1.02	0.20	0.51	0.50	73.0	71.1
Lec_1_11	249	116	0.466	0.659	0.524	1.02	0.15	0.98	-0.40	0.98	-0.10	0.52	0.51	73.0	71.3
Lec_1_12	249	157	0.631	0.498	0.400	0.13	0.15	1.12	1.70	1.18	1.50	0.40	0.48	69.7	72.7
Lec_1_13	249	131	0.526	0.643	0.482	0.70	0.15	1.03	0.60	1.00	0.10	0.48	0.50	66.8	71.1
Lec_1_14	249	176	0.707	0.643	0.517	-0.32	0.16	0.93	-0.90	0.84	-1.10	0.50	0.45	78.3	75.7
Lec_1_15	249	131	0.526	0.723	0.501	0.70	0.15	0.99	-0.10	1.05	0.60	0.50	0.50	70.9	71.1
Lec_2_1	218	146	0.670	0.404	0.388	-0.41	0.16	1.09	1.30	1.20	1.50	0.38	0.46	70.0	73.8
Lec_2_2	218	155	0.711	0.532	0.492	-0.66	0.17	0.96	-0.50	0.85	-1.00	0.48	0.44	73.3	75.5
Lec_2_3	218	160	0.737	0.516	0.481	-0.80	0.17	0.95	-0.60	0.86	-0.80	0.47	0.43	75.6	76.8
Lec_2_4	218	174	0.798	0.624	0.630	-1.24	0.19	0.73	-2.90	0.47	-3.00	0.60	0.39	84.8	81.0
Lec_2_5	218	100	0.459	0.550	0.470	0.73	0.16	1.04	0.60	1.02	0.20	0.47	0.49	72.4	72.0
Lec_2_6	218	171	0.784	0.257	0.359	-1.14	0.18	1.05	0.60	1.07	0.40	0.36	0.40	81.6	80.1
Lec_2_7	218	186	0.853	0.220	0.273	-1.70	0.21	1.09	0.70	1.38	1.30	0.26	0.35	83.9	85.6
Lec_2_8	218	75	0.344	0.606	0.511	1.37	0.16	0.93	-1.00	0.91	-0.70	0.53	0.48	76.0	74.3
Lec_2_9	218	43	0.197	0.294	0.268	2.34	0.19	1.10	0.90	1.87	3.20	0.30	0.42	82.0	81.8
Lec_2_10	218	123	0.564	0.679	0.553	0.17	0.16	0.92	-1.30	0.87	-1.30	0.55	0.49	77.4	71.8
Lec_2_11	218	101	0.463	0.569	0.456	0.71	0.16	1.05	0.70	1.10	1.00	0.45	0.49	71.0	71.9

Lec_2_12	218	91	0.417	0.771	0.563	0.96	0.16	0.88	-1.70	0.90	-0.90	0.57	0.49	77.0	72.6
Lec_2_13	218	103	0.472	0.404	0.332	0.66	0.16	1.23	3.20	1.29	2.80	0.33	0.49	61.8	71.9
Lec_2_14	218	138	0.636	0.627	0.500	-0.20	0.16	0.97	-0.40	0.91	-0.70	0.49	0.47	74.7	72.5
Lec_2_15	218	159	0.729	0.532	0.511	-0.77	0.17	0.92	-1.00	0.97	-0.10	0.49	0.43	77.9	76.5
Exam_1	305	229	0.746	0.495	0.509	-1.99	0.15	0.90	-1.20	0.76	-1.50	0.52	0.44	81.5	79.4
Exam_2	305	129	0.420	0.638	0.563	-0.15	0.13	0.90	-2.00	0.88	-1.40	0.54	0.47	76.1	70.3
Exam_3	305	63	0.206	0.353	0.374	1.18	0.16	1.06	0.70	1.17	1.10	0.37	0.43	82.5	81.9
Exam_4	305	48	0.156	0.430	0.513	1.59	0.18	0.86	-1.30	0.71	-1.50	0.50	0.40	86.9	85.6
Exam_5	305	95	0.309	0.391	0.389	0.47	0.14	1.10	1.50	1.20	1.80	0.38	0.46	70.4	74.7
Exam_6	305	176	0.573	0.469	0.379	-0.96	0.13	1.13	2.40	1.28	2.80	0.38	0.47	64.0	69.9
Exam_7	305	150	0.489	0.404	0.370	-0.51	0.13	1.14	2.70	1.27	3.10	0.37	0.48	65.3	69.3
Exam_8	305	94	0.306	0.508	0.456	0.49	0.14	1.02	0.30	1.05	0.50	0.44	0.46	74.7	74.8
Exam_9	305	114	0.371	0.612	0.535	0.11	0.13	0.93	-1.30	0.90	-1.10	0.52	0.47	73.4	71.6
Exam_10	305	133	0.433	0.678	0.570	-0.22	0.13	0.89	-2.20	0.80	-2.60	0.56	0.48	74.1	70.1
Red_Exam_1	286	277	0.969	0.126	0.250	-3.00	0.35	0.96	0.00	0.66	-0.50	0.20	0.16	96.8	96.8
Red_Exam_2	286	267	0.934	0.154	0.218	-2.18	0.25	1.02	0.10	3.98	4.60	0.16	0.23	93.2	93.2
Red_Exam_3	286	210	0.734	0.448	0.360	-0.30	0.15	1.10	1.40	1.24	1.50	0.33	0.41	75.4	76.2
Red_Exam_4	286	255	0.892	0.294	0.355	-1.59	0.20	0.97	-0.20	0.80	-0.60	0.31	0.29	89.0	89.0
Red_Exam_5	286	180	0.629	0.531	0.456	0.32	0.14	1.04	0.70	1.13	1.20	0.43	0.46	70.5	72.3
Red_Exam_6	286	178	0.622	0.713	0.527	0.35	0.14	0.95	-0.80	0.97	-0.20	0.49	0.46	75.4	72.1
Red_Exam_7	286	194	0.678	0.573	0.462	0.04	0.14	1.00	0.00	1.11	0.90	0.43	0.44	73.7	73.7
Red_Exam_8	286	236	0.825	0.490	0.494	-0.95	0.17	0.90	-1.10	0.73	-1.20	0.43	0.35	83.3	82.6
Red_Exam_9	286	65	0.227	0.378	0.341	2.66	0.16	1.14	1.50	1.34	1.90	0.38	0.49	79.7	81.3
Red_Exam_10	286	163	0.570	0.573	0.421	0.64	0.14	1.09	1.60	1.06	0.70	0.42	0.48	66.2	71.3
Red_Exam_11	286	154	0.538	0.657	0.478	0.81	0.14	1.00	0.10	1.00	0.10	0.48	0.49	71.9	71.1
Red_Exam_12	286	175	0.612	0.797	0.594	0.41	0.14	0.85	-2.60	0.74	-2.60	0.57	0.47	76.2	71.9
Red_Exam_13	286	138	0.483	0.643	0.495	1.11	0.14	1.00	0.10	0.98	-0.20	0.50	0.50	70.1	71.3
Red_Exam_14	286	196	0.685	0.657	0.508	0.00	0.14	0.95	-0.70	0.94	-0.40	0.47	0.44	75.1	73.9
Red_Exam_15	286	84	0.294	0.322	0.247	2.19	0.15	1.24	3.00	1.67	4.30	0.30	0.50	74.0	77.2
Red_Exam_16	287	187	0.652	0.697	0.526	0.20	0.14	0.92	-1.40	0.84	-1.40	0.51	0.45	75.9	72.8

Red_Exam_17	287	214	0.746	0.641	0.525	-0.37	0.15	0.88	-1.70	0.79	-1.30	0.49	0.41	80.5	76.8
Red_Exam_18	287	234	0.815	0.432	0.447	-0.86	0.17	0.91	-1.00	1.46	2.00	0.39	0.36	84.0	81.8
Red_Exam_19	287	164	0.571	0.808	0.582	0.64	0.14	0.85	-2.70	0.78	-2.50	0.58	0.48	77.3	71.4
Red_Exam_20	287	218	0.760	0.362	0.324	-0.46	0.15	1.11	1.50	1.19	1.10	0.32	0.40	73.4	77.6
Red_Exam_21	287	211	0.735	0.627	0.531	-0.30	0.15	0.88	-1.70	0.76	-1.60	0.50	0.41	79.8	76.3
Red_Exam_22	287	241	0.840	0.348	0.333	-1.06	0.17	1.03	0.30	1.08	0.40	0.31	0.34	83.3	84.0
Red_Exam_23	**Unused**														
Red_Exam_24															
Red_Exam_25	287	113	0.394	0.697	0.535	1.60	0.14	0.93	-1.10	0.92	-0.80	0.55	0.50	73.8	72.9
Red_Exam_26	287	219	0.766	0.476	0.420	-0.48	0.15	1.00	0.10	1.12	0.70	0.40	0.40	77.3	77.9
Red_Exam_27	287	104	0.362	0.711	0.548	1.78	0.14	0.93	-1.10	0.83	-1.60	0.56	0.50	75.5	74.1
Red_Exam_28	287	187	0.652	0.697	0.511	0.20	0.14	0.94	-1.00	0.91	-0.80	0.49	0.45	74.5	72.8
Red_Exam_29	287	203	0.707	0.307	0.296	-0.12	0.15	1.20	2.90	1.27	1.80	0.29	0.43	69.1	75.0
Red_Exam_30	287	248	0.864	0.293	0.311	-1.29	0.18	1.02	0.30	1.10	0.50	0.29	0.32	86.5	86.3

Table 63: Breakdown of Individual Items used in Foundations of Chemistry IB 2014 Multiple-Choice Assessments using CTT and Rasch Analysis

Foundations of Chemistry IB 2014															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	216	183	0.847	0.204	0.328	-1.38	0.21	1.12	0.90	1.37	1.30	0.31	0.39	82.9	85.6
Lec_1_2	216	46	0.213	0.333	0.363	2.67	0.19	1.18	1.60	1.18	0.70	0.38	0.47	77.1	82.5
Lec_1_3	216	189	0.875	0.148	0.323	-1.67	0.23	1.02	0.20	1.34	1.00	0.32	0.37	89.5	88.0
Lec_1_4	216	161	0.745	0.537	0.594	-0.57	0.18	0.84	-1.90	0.65	-2.20	0.57	0.46	81.0	78.3
Lec_1_5	216	163	0.755	0.500	0.546	-0.64	0.18	0.89	-1.20	0.79	-1.20	0.52	0.45	81.9	78.9
Lec_1_6	216	190	0.880	0.241	0.497	-1.72	0.23	0.83	-1.10	0.57	-1.40	0.47	0.36	91.0	88.4
Lec_1_7	216	112	0.519	0.722	0.631	0.77	0.16	0.83	-2.60	0.75	-2.40	0.62	0.52	78.6	71.8
Lec_1_8	216	168	0.778	0.315	0.378	-0.80	0.19	1.09	1.00	1.26	1.20	0.37	0.44	79.5	80.4
Lec_1_9	216	95	0.440	0.556	0.481	1.20	0.16	1.02	0.20	1.10	0.80	0.50	0.52	70.5	72.7
Lec_1_10	216	126	0.583	0.593	0.556	0.41	0.16	0.94	-0.80	0.86	-1.30	0.55	0.51	71.9	72.2
Lec_1_11	216	112	0.519	0.648	0.590	0.77	0.16	0.89	-1.60	0.82	-1.70	0.59	0.52	77.6	71.8
Lec_1_12	216	150	0.698	0.298	0.349	-0.24	0.17	1.22	2.60	1.15	1.00	0.36	0.48	67.1	75.4
Lec_1_13	216	111	0.514	0.500	0.468	0.79	0.16	1.06	0.90	1.11	1.00	0.48	0.52	68.6	71.8
Lec_1_14	216	161	0.745	0.407	0.499	-0.57	0.18	0.97	-0.40	0.94	-0.30	0.48	0.46	78.1	78.3
Lec_1_15	216	103	0.477	0.574	0.464	1.00	0.16	1.09	1.30	1.17	1.50	0.46	0.52	66.7	72.2
Lec_2_1	184	134	0.728	0.674	0.435	-0.43	0.19	1.05	0.60	0.99	0.00	0.42	0.45	73.9	76.5
Lec_2_2	184	137	0.745	0.761	0.516	-0.54	0.19	0.92	-0.80	0.81	-0.90	0.49	0.44	80.1	77.5
Lec_2_3	184	137	0.745	0.717	0.475	-0.54	0.19	0.97	-0.20	0.95	-0.20	0.45	0.44	77.8	77.5
Lec_2_4	184	140	0.761	0.761	0.584	-0.65	0.20	0.81	-2.00	0.70	-1.50	0.54	0.43	81.8	78.8
Lec_2_5	184	88	0.478	0.913	0.621	1.01	0.17	0.85	-2.20	0.79	-1.80	0.62	0.53	76.1	72.1
Lec_2_6	184	151	0.821	0.609	0.423	-1.11	0.21	0.99	0.00	0.83	-0.60	0.41	0.39	83.5	83.2
Lec_2_7	184	168	0.913	0.435	0.362	-2.10	0.28	0.94	-0.20	0.70	-0.60	0.34	0.30	92.0	91.1
Lec_2_8	184	48	0.261	0.543	0.427	2.39	0.20	1.17	1.50	1.17	0.80	0.45	0.54	75.0	80.7
Lec_2_9	184	52	0.283	0.609	0.460	2.23	0.20	1.07	0.70	1.16	0.80	0.49	0.54	79.5	79.6
Lec_2_10	184	133	0.727	0.612	0.403	-0.39	0.19	1.08	0.90	1.12	0.70	0.40	0.45	76.7	76.2
Lec_2_11	184	176	0.957	0.239	0.097	-2.94	0.38	1.17	0.60	2.28	1.70	0.10	0.23	95.5	95.4

Lec_2_12	184	104	0.565	0.783	0.471	0.53	0.17	1.07	1.10	1.02	0.20	0.47	0.51	64.8	70.6
Lec_2_13	184	77	0.418	0.891	0.628	1.36	0.18	0.85	-1.90	0.75	-2.00	0.63	0.54	80.1	73.6
Lec_2_14	184	83	0.451	0.848	0.529	1.17	0.18	0.99	-0.10	1.06	0.50	0.53	0.53	75.6	72.9
Lec_2_15	184	121	0.658	0.609	0.395	0.01	0.18	1.14	1.90	1.11	0.80	0.40	0.48	65.9	72.6
Exam_1	276	216	0.783	0.304	0.314	-1.55	0.16	1.09	1.10	1.00	0.10	0.31	0.37	77.4	79.5
Exam_2	276	118	0.428	0.638	0.493	0.40	0.14	0.97	-0.50	1.00	0.00	0.49	0.47	74.1	70.5
Exam_3	276	65	0.236	0.435	0.425	1.56	0.16	1.03	0.40	1.07	0.50	0.44	0.47	80.7	80.7
Exam_4	276	68	0.246	0.507	0.501	1.48	0.16	0.93	-0.80	0.93	-0.40	0.51	0.47	81.9	80.0
Exam_5	276	73	0.264	0.435	0.386	1.35	0.16	1.08	1.00	1.39	2.50	0.39	0.47	76.3	78.7
Exam_6	276	135	0.489	0.623	0.480	0.08	0.14	0.98	-0.30	0.98	-0.20	0.47	0.46	69.3	68.7
Exam_7	276	167	0.605	0.507	0.441	-0.51	0.14	0.98	-0.30	1.13	1.40	0.43	0.44	74.4	69.4
Exam_8	276	138	0.500	0.725	0.566	0.03	0.14	0.88	-2.50	0.81	-2.60	0.56	0.46	73.3	68.4
Exam_9	276	221	0.801	0.362	0.322	-1.68	0.16	1.00	0.00	1.59	2.90	0.32	0.35	81.5	80.8
Exam_10	276	199	0.721	0.507	0.423	-1.16	0.15	0.96	-0.60	1.05	0.40	0.42	0.40	76.3	74.5
Red_Exam_1	259	253	0.917	0.174	0.502	-3.69	0.42	1.05	0.30	6.23	4.60	-0.06	0.13	97.7	97.7
Red_Exam_2	259	238	0.862	0.290	0.509	-2.31	0.24	1.04	0.30	1.12	0.40	0.19	0.23	91.9	91.9
Red_Exam_3	259	190	0.688	0.551	0.543	-0.71	0.15	0.97	-0.40	0.87	-0.80	0.41	0.38	76.0	75.5
Red_Exam_4	259	219	0.793	0.522	0.602	-1.51	0.18	0.89	-1.00	0.67	-1.40	0.40	0.31	86.0	84.7
Red_Exam_5	259	162	0.587	0.681	0.548	-0.11	0.14	0.95	-0.90	0.88	-1.10	0.47	0.42	69.8	70.3
Red_Exam_6	259	179	0.649	0.652	0.549	-0.46	0.15	0.96	-0.70	0.87	-0.90	0.44	0.40	74.4	73.0
Red_Exam_7	259	180	0.652	0.580	0.506	-0.49	0.15	1.02	0.30	1.04	0.30	0.38	0.40	73.3	73.2
Red_Exam_8	259	204	0.739	0.522	0.569	-1.06	0.16	0.94	-0.70	0.79	-1.10	0.41	0.35	81.4	79.6
Red_Exam_9	259	64	0.232	0.319	0.325	1.97	0.16	1.21	2.30	1.25	1.60	0.32	0.46	74.0	80.0
Red_Exam_10	259	132	0.478	0.652	0.523	0.49	0.14	0.97	-0.60	0.89	-1.20	0.49	0.45	69.8	69.4
Red_Exam_11	259	129	0.469	0.669	0.513	0.55	0.14	0.97	-0.50	0.96	-0.40	0.48	0.46	70.9	69.5
Red_Exam_12	259	141	0.511	0.696	0.542	0.31	0.14	0.95	-0.90	0.85	-1.60	0.50	0.45	66.3	69.2
Red_Exam_13	259	126	0.457	0.681	0.550	0.61	0.14	0.93	-1.40	0.89	-1.30	0.52	0.46	70.9	69.6
Red_Exam_14	259	184	0.667	0.594	0.530	-0.57	0.15	1.00	0.00	0.88	-0.80	0.40	0.39	71.7	74.1
Red_Exam_15	259	64	0.232	0.377	0.372	1.97	0.16	1.04	0.50	1.40	2.50	0.40	0.46	79.5	80.0
Red_Exam_16	259	167	0.605	0.609	0.490	-0.21	0.14	1.05	0.80	0.96	-0.30	0.39	0.42	70.9	70.9

Red_Exam_17	259	175	0.634	0.768	0.641	-0.38	0.15	0.80	-3.50	0.80	-1.60	0.54	0.40	79.8	72.2
Red_Exam_18	259	200	0.725	0.652	0.633	-0.96	0.16	0.83	-2.20	0.68	-1.90	0.49	0.36	80.6	78.4
Red_Exam_19	259	120	0.435	0.725	0.548	0.73	0.14	0.91	-1.60	0.86	-1.60	0.53	0.46	73.3	70.0
Red_Exam_20	259	182	0.662	0.495	0.461	-0.53	0.15	1.08	1.20	1.10	0.80	0.33	0.39	71.7	73.6
Red_Exam_21	259	196	0.710	0.609	0.587	-0.86	0.16	0.91	-1.20	0.76	-1.40	0.44	0.36	76.7	77.2
Red_Exam_22	259	219	0.793	0.522	0.555	-1.51	0.18	0.96	-0.40	0.80	-0.80	0.35	0.31	86.0	84.7
Red_Exam_23	259	62	0.225	0.377	0.314	2.03	0.17	1.22	2.30	1.33	2.00	0.30	0.46	74.8	80.5
Red_Exam_24	259	67	0.243	0.493	0.426	1.90	0.16	1.01	0.20	1.00	0.10	0.45	0.46	79.8	79.3
Red_Exam_25	259	91	0.330	0.464	0.378	1.33	0.15	1.17	2.30	1.18	1.70	0.35	0.47	69.4	74.1
Red_Exam_26	259	124	0.449	0.348	0.339	0.65	0.14	1.25	4.10	1.29	3.10	0.28	0.46	61.6	69.8
Red_Exam_27	259	127	0.460	0.536	0.449	0.59	0.14	1.08	1.40	1.19	2.10	0.39	0.46	67.4	69.6
Red_Exam_28	259	167	0.605	0.652	0.553	-0.21	0.14	0.95	-0.90	0.88	-1.00	0.46	0.42	72.5	70.9
Red_Exam_29	259	111	0.402	0.652	0.520	0.91	0.14	0.94	-1.00	0.94	-0.60	0.51	0.47	74.8	70.9
Red_Exam_30	259	81	0.293	0.478	0.439	1.55	0.15	1.03	0.40	1.00	0.10	0.45	0.47	75.2	76.1

Table 64: Breakdown of Individual Items used in Foundations of Chemistry IB 2015 Multiple-Choice Assessments using CTT and Rasch Analysis

Foundations of Chemistry IB 2015															
	Counts		Classical Test Theory			Rasch Analysis									
Item	Count	Score	P	D	r_{pbi}	Item Difficulty	Model S.E.	Infit	ZSTD	Outfit	ZSTD	Obs. Correl.	Exp. Correl.	OBS%	EXP%
Lec_1_1	231	203	0.879	0.416	0.404	-1.83	0.22	0.88	-0.80	1.51	1.50	0.38	0.33	88.6	88.2
Lec_1_2	231	48	0.208	0.416	0.418	2.42	0.19	1.09	0.80	0.90	-0.40	0.42	0.46	79.8	82.4
Lec_1_3	231	200	0.866	0.398	0.303	-1.70	0.21	1.01	0.20	1.29	1.00	0.30	0.34	87.3	87.0
Lec_1_4	231	170	0.736	0.623	0.436	-0.68	0.17	1.00	0.00	1.00	0.10	0.42	0.42	74.1	76.8
Lec_1_5	231	164	0.710	0.641	0.489	-0.52	0.16	0.94	-0.70	0.89	-0.70	0.47	0.44	75.9	75.1
Lec_1_6	231	193	0.835	0.606	0.506	-1.41	0.19	0.87	-1.10	0.60	-1.80	0.47	0.37	83.3	84.2
Lec_1_7	231	88	0.381	0.727	0.589	1.29	0.16	0.87	-1.90	0.75	-2.20	0.59	0.49	73.2	73.1
Lec_1_8	231	173	0.749	0.519	0.393	-0.77	0.17	1.05	0.60	0.95	-0.20	0.39	0.42	77.2	77.7
Lec_1_9	231	81	0.351	0.589	0.487	1.46	0.16	0.98	-0.20	1.02	0.20	0.50	0.49	73.7	74.6
Lec_1_10	231	124	0.539	0.626	0.425	0.45	0.15	1.08	1.30	1.11	1.00	0.43	0.48	68.0	71.0
Lec_1_11	231	113	0.489	0.710	0.559	0.70	0.15	0.91	-1.50	0.87	-1.30	0.55	0.49	78.1	71.3
Lec_1_12	231	139	0.602	0.745	0.472	0.10	0.15	1.01	0.20	1.07	0.60	0.46	0.47	67.5	71.2
Lec_1_13	231	120	0.519	0.502	0.391	0.54	0.15	1.13	2.00	1.26	2.40	0.39	0.49	68.0	71.0
Lec_1_14	231	166	0.719	0.623	0.493	-0.57	0.16	0.95	-0.70	0.83	-1.10	0.48	0.43	76.8	75.6
Lec_1_15	231	120	0.519	0.589	0.365	0.54	0.15	1.16	2.50	1.18	1.70	0.38	0.49	64.5	71.0
Lec_2_1	198	144	0.727	0.525	0.518	-0.48	0.18	0.92	-1.00	0.76	-1.40	0.50	0.43	80.9	76.1
Lec_2_2	198	159	0.803	0.424	0.446	-1.00	0.20	0.95	-0.40	0.85	-0.60	0.42	0.38	83.0	81.2
Lec_2_3	198	140	0.707	0.566	0.508	-0.35	0.18	0.93	-0.80	0.83	-1.00	0.49	0.44	74.7	75.2
Lec_2_4	198	148	0.747	0.586	0.548	-0.61	0.18	0.87	-1.60	0.68	-1.80	0.52	0.42	79.4	77.3
Lec_2_5	198	86	0.434	0.667	0.492	1.16	0.17	0.99	-0.20	1.11	1.00	0.50	0.51	75.3	72.3
Lec_2_6	198	154	0.778	0.404	0.426	-0.82	0.19	0.98	-0.20	0.96	-0.10	0.41	0.40	80.4	79.4
Lec_2_7	198	175	0.884	0.222	0.376	-1.74	0.24	0.94	-0.30	0.77	-0.60	0.35	0.31	87.6	88.4
Lec_2_8	198	62	0.313	0.485	0.350	1.87	0.18	1.20	2.10	1.48	2.70	0.36	0.51	71.1	76.8
Lec_2_9	198	47	0.237	0.505	0.486	2.40	0.20	0.95	-0.40	0.91	-0.40	0.52	0.49	84.0	81.1
Lec_2_10	198	135	0.682	0.465	0.448	-0.20	0.17	1.02	0.30	1.00	0.00	0.44	0.45	75.3	73.9
Lec_2_11	198	189	0.955	-0.040	0.161	-2.86	0.35	0.96	0.00	2.12	1.70	0.16	0.21	95.4	95.4

Lec_2_12	198	119	0.604	0.548	0.440	0.26	0.17	1.07	1.00	1.03	0.30	0.44	0.48	68.6	71.7
Lec_2_13	198	68	0.343	0.707	0.556	1.69	0.18	0.90	-1.20	0.89	-0.80	0.57	0.51	78.4	75.7
Lec_2_14	198	92	0.465	0.505	0.401	0.99	0.17	1.18	2.40	1.21	1.90	0.40	0.51	63.9	71.3
Lec_2_15	198	138	0.697	0.525	0.453	-0.29	0.17	1.02	0.30	0.95	-0.30	0.44	0.44	72.7	74.7
Exam_1	300	231	0.770	0.520	0.319	-1.36	0.15	1.09	1.20	1.10	0.70	0.32	0.38	77.0	78.6
Exam_2	300	151	0.505	0.856	0.542	0.12	0.13	0.90	-1.90	0.90	-1.30	0.53	0.46	73.6	69.4
Exam_3	300	79	0.264	0.495	0.381	1.46	0.15	1.07	1.00	1.10	0.80	0.39	0.45	77.7	78.1
Exam_4	300	72	0.240	0.440	0.364	1.61	0.15	1.05	0.60	1.23	1.50	0.38	0.44	79.4	79.6
Exam_5	300	90	0.300	0.640	0.455	1.22	0.14	1.00	0.10	1.01	0.10	0.45	0.45	73.3	75.7
Exam_6	300	150	0.500	0.693	0.472	0.14	0.13	0.99	-0.20	0.98	-0.20	0.47	0.46	69.9	69.5
Exam_7	300	182	0.607	0.773	0.498	-0.41	0.13	0.95	-1.00	0.89	-1.20	0.49	0.44	72.0	70.5
Exam_8	300	150	0.500	0.853	0.567	0.14	0.13	0.88	-2.50	0.83	-2.30	0.56	0.46	75.3	69.5
Exam_9	300	251	0.839	0.455	0.300	-1.87	0.17	1.02	0.20	1.23	1.10	0.30	0.35	84.1	84.1
Exam_10	300	217	0.723	0.560	0.384	-1.06	0.14	1.01	0.20	1.16	1.20	0.38	0.41	73.3	75.0
Red_Exam_1	277	266	0.887	0.360	0.613	-3.06	0.32	0.99	0.00	2.37	2.30	0.15	0.19	96.0	96.0
Red_Exam_2	277	252	0.840	0.373	0.505	-2.11	0.22	1.14	0.90	1.78	2.10	0.11	0.27	91.0	91.0
Red_Exam_3	277	202	0.673	0.693	0.569	-0.58	0.15	0.97	-0.50	1.03	0.20	0.41	0.40	78.3	76.2
Red_Exam_4	277	237	0.790	0.653	0.665	-1.51	0.18	0.86	-1.30	0.66	-1.50	0.44	0.32	86.6	85.9
Red_Exam_5	277	176	0.587	0.707	0.540	-0.05	0.14	0.99	-0.10	0.99	0.00	0.43	0.43	72.9	71.6
Red_Exam_6	277	176	0.587	0.760	0.556	-0.05	0.14	0.97	-0.50	0.90	-1.00	0.46	0.43	72.2	71.6
Red_Exam_7	277	190	0.633	0.773	0.582	-0.33	0.14	0.95	-0.70	0.84	-1.30	0.47	0.41	73.3	73.7
Red_Exam_8	277	221	0.737	0.640	0.587	-1.04	0.16	0.96	-0.40	0.90	-0.50	0.39	0.36	81.9	81.1
Red_Exam_9	277	62	0.207	0.387	0.338	2.20	0.16	1.05	0.60	1.24	1.50	0.35	0.41	80.5	80.4
Red_Exam_10	277	177	0.590	0.773	0.552	-0.07	0.14	0.98	-0.30	0.91	-0.90	0.45	0.43	73.3	71.7
Red_Exam_11	277	139	0.463	0.547	0.455	0.63	0.13	1.07	1.40	1.13	1.60	0.38	0.45	69.3	69.1
Red_Exam_12	277	147	0.490	0.813	0.581	0.49	0.13	0.88	-2.30	0.87	-1.50	0.53	0.44	73.3	69.2
Red_Exam_13	277	153	0.510	0.667	0.511	0.38	0.14	1.00	0.10	0.97	-0.30	0.44	0.44	70.4	69.4
Red_Exam_14	277	169	0.563	0.813	0.578	0.08	0.14	0.93	-1.40	0.84	-1.70	0.50	0.43	74.7	70.7
Red_Exam_15	277	76	0.253	0.307	0.274	1.86	0.15	1.16	2.00	1.59	3.90	0.25	0.43	76.5	76.9
Red_Exam_16	277	199	0.663	0.653	0.562	-0.52	0.15	1.00	0.00	0.89	-0.80	0.42	0.40	74.4	75.5

Red_Exam_17	277	189	0.630	0.880	0.690	-0.31	0.14	0.77	-4.00	0.66	-3.20	0.60	0.41	80.9	73.5
Red_Exam_18	277	214	0.713	0.707	0.632	-0.86	0.16	0.88	-1.40	0.84	-1.00	0.47	0.38	81.6	79.2
Red_Exam_19	277	153	0.510	0.853	0.619	0.38	0.14	0.83	-3.40	0.76	-3.00	0.58	0.44	76.2	69.4
Red_Exam_20	277	205	0.683	0.427	0.437	-0.65	0.15	1.19	2.40	1.69	3.90	0.22	0.39	72.9	76.9
Red_Exam_21	277	198	0.662	0.709	0.605	-0.50	0.15	0.94	-0.90	0.78	-1.70	0.47	0.40	72.6	75.2
Red_Exam_22	277	227	0.757	0.613	0.619	-1.20	0.17	0.92	-0.80	0.80	-1.00	0.42	0.35	84.1	82.8
Red_Exam_23	277	68	0.227	0.413	0.373	2.05	0.15	1.03	0.40	1.02	0.20	0.39	0.42	78.3	78.8
Red_Exam_24	277	122	0.407	0.600	0.415	0.94	0.14	1.12	2.10	1.17	2.00	0.35	0.45	65.3	69.6
Red_Exam_25	277	119	0.397	0.507	0.363	1.00	0.14	1.19	3.40	1.24	2.60	0.30	0.45	61.4	69.8
Red_Exam_26	277	162	0.540	0.533	0.451	0.22	0.14	1.12	2.10	1.08	0.90	0.35	0.44	63.9	70.0
Red_Exam_27	277	152	0.508	0.722	0.509	0.40	0.14	1.01	0.20	0.99	-0.10	0.44	0.44	70.0	69.4
Red_Exam_28	277	187	0.623	0.707	0.531	-0.27	0.14	1.02	0.40	1.03	0.30	0.40	0.42	73.6	73.2
Red_Exam_29	277	124	0.413	0.840	0.555	0.91	0.14	0.87	-2.50	0.87	-1.60	0.55	0.45	76.9	69.4
Red_Exam_30	277	90	0.300	0.453	0.367	1.56	0.14	1.12	1.80	1.16	1.40	0.35	0.44	70.4	73.9

7.6 MCQ Assessment Student Ability and Item Difficulty Rasch Measures Test for Normality

Table 65: Tests of Normality on Rasch Student Ability Measures from Chemistry IA MCQ Assessment Tasks

			Rasch Measures				Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean	Standard Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Chemistry IA Students	2012	Lecture Test 1	469	0.508	1.251	0.058	0.340	0.113	0.222	0.225	0.086	469	<<0.001	0.976	469	<<0.001
		Lecture Test 2	446	0.312	0.965	0.046	-0.054	0.116	-0.133	0.231	0.083	446	<<0.001	0.983	446	<<0.001
		Exam Test	508	-0.208	1.162	0.052	0.177	0.108	0.539	0.216	0.105	508	<<0.001	0.971	508	<<0.001
		Redeemable Exam	487	0.509	1.099	0.050	0.402	0.111	0.033	0.221	0.070	487	<<0.001	0.984	487	<<0.001
	2013	Lecture Test 1	448	0.640	1.269	0.060	0.278	0.115	-0.023	0.230	0.076	448	<<0.001	0.978	448	<<0.001
		Lecture Test 2	420	0.479	1.027	0.050	-0.112	0.119	-0.036	0.238	0.093	420	<<0.001	0.978	420	<<0.001
		Exam Test	505	-0.136	1.119	0.050	0.132	0.109	0.534	0.217	0.101	505	<<0.001	0.970	505	<<0.001
		Redeemable Exam	488	0.502	1.111	0.050	0.355	0.111	0.134	0.221	0.077	488	<<0.001	0.987	488	<<0.001
	2014	Lecture Test 1	474	0.322	1.172	0.054	0.364	0.112	0.626	0.224	0.097	474	<<0.001	0.975	474	<<0.001
		Lecture Test 2	436	0.536	1.040	0.050	-0.294	0.117	0.575	0.233	0.092	436	<<0.001	0.978	436	<<0.001
		Exam Test	508	0.570	1.262	0.056	0.218	0.108	-0.402	2.160	0.112	508	<<0.001	0.966	508	<<0.001
		Redeemable Exam	499	0.491	1.064	0.048	0.376	0.109	0.666	0.218	0.068	499	<<0.001	0.986	499	<<0.001
	2015	Lecture Test 1	504	0.471	1.180	0.053	0.488	0.109	0.621	0.217	0.117	504	<<0.001	0.970	504	<<0.001
		Lecture Test 2	451	0.740	1.119	0.053	0.470	0.115	-0.001	0.229	0.117	451	<<0.001	0.966	451	<<0.001
		Exam Test	546	0.466	1.241	0.053	-0.078	0.105	0.263	0.209	0.093	546	<<0.001	0.973	546	<<0.001
		Redeemable Exam	525	0.432	0.957	0.042	0.123	0.107	0.659	0.213	0.071	525	<<0.001	0.989	525	0.001

Table 66: Tests of Normality on Rasch Item Difficulty Measures from Chemistry IA MCQ Assessment Tasks

			Rasch Measures				Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean	Standard Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Chemistry IA Items	2012	Lecture Test 1	15	0.001	0.850	0.219	-0.122	0.580	-1.153	1.121	0.184	15	0.186	0.936	15	0.333
		Lecture Test 2	15	-0.001	1.000	0.258	1.344	0.580	1.294	1.121	0.312	15	<<0.001	0.839	15	0.012
		Exam Test	10	0.000	0.849	0.268	0.249	0.687	-0.605	1.334	0.180	10	0.200	0.928	10	0.427
		Redeemable Exam	30	-0.001	0.975	0.178	1.009	0.427	1.906	0.833	0.147	30	0.095	0.941	30	0.099
		Stacked	70	0.000	0.893	0.107	0.715	0.287	0.585	0.566	0.091	70	0.200	0.963	70	0.035
	2013	Lecture Test 1	15	0.001	0.845	0.218	-0.090	0.580	-1.106	1.121	0.180	15	0.200	0.929	15	0.260
		Lecture Test 2	15	-0.001	1.138	0.294	1.482	0.580	1.701	1.121	0.268	15	0.005	0.819	15	0.007
		Exam Test	10	-0.001	0.803	0.254	0.279	0.687	0.145	1.334	0.177	10	0.200	0.965	10	0.842
		Redeemable Exam	30	0.000	0.955	0.174	0.808	0.427	0.985	0.833	0.111	30	0.200	0.960	30	0.303
		Stacked	70	0.000	0.892	0.107	0.834	0.287	0.765	0.566	0.081	70	0.200	0.956	70	0.015
	2014	Lecture Test 1	15	-0.001	1.181	0.305	0.427	0.580	-0.890	1.121	0.160	15	0.200	0.938	15	0.362
		Lecture Test 2	15	-0.001	0.809	0.209	0.923	0.580	0.664	1.121	0.135	15	0.200	0.926	15	0.242
		Exam Test	10	0.000	0.816	0.258	0.556	0.687	-1.187	1.334	0.206	10	0.200	0.901	10	0.223
		Redeemable Exam	30	0.000	0.866	0.158	0.255	0.427	-0.609	0.833	0.084	30	0.200	0.967	30	0.457
		Stacked	70	0.000	0.843	0.101	0.525	0.287	-0.418	0.566	0.099	70	0.087	0.961	70	0.030
	2015	Lecture Test 1	15	-0.001	1.164	0.301	0.453	0.580	-1.067	1.121	0.118	15	0.200	0.924	15	0.221
		Lecture Test 2	15	-0.001	0.772	0.199	1.434	0.580	2.798	1.121	0.171	15	0.200	0.896	15	0.083
		Exam Test	10	0.000	0.788	0.249	0.427	0.687	-1.645	1.334	0.190	10	0.200	0.887	10	0.156
		Redeemable Exam	30	-0.001	0.957	0.175	0.719	0.427	0.178	0.833	0.125	30	0.200	0.953	30	0.198
		Stacked	70	0.001	0.851	0.102	0.739	0.287	-0.035	0.566	0.118	70	0.017	0.947	70	0.005

Table 67: Tests of Normality on Rasch Student Ability Measures from Chemistry IB MCQ Assessment Tasks

			Rasch Measures				Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean	Standard Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Chemistry IB Students	2012	Lecture Test 1	382	0.536	1.090	0.056	0.713	0.125	0.272	0.249	0.129	382	<<0.001	0.949	382	<<0.001
		Lecture Test 2	364	-0.015	1.198	0.063	0.617	0.128	1.163	0.255	0.110	364	<<0.001	0.954	364	<<0.001
		Exam Test	433	-0.171	1.147	0.055	0.043	0.117	0.382	0.234	0.112	433	<<0.001	0.969	433	<<0.001
		Redeemable Exam	421	0.456	0.982	0.048	0.600	0.119	0.635	0.237	0.086	421	<<0.001	0.973	421	<<0.001
	2013	Lecture Test 1	378	0.649	1.101	0.057	0.552	0.125	0.184	0.250	0.106	378	<<0.001	0.965	378	<<0.001
		Lecture Test 2	348	0.146	1.098	0.059	0.273	0.131	0.332	0.261	0.100	348	<<0.001	0.981	348	<<0.001
		Exam Test	450	0.116	1.209	0.057	-0.026	0.115	0.184	0.230	0.096	450	<<0.001	0.975	450	<<0.001
		Redeemable Exam	434	0.646	1.107	0.053	0.536	0.117	0.452	0.234	0.096	434	<<0.001	0.976	434	<<0.001
	2014	Lecture Test 1	423	0.709	1.242	0.060	0.422	0.119	-0.161	0.237	0.110	423	<<0.001	0.967	423	<<0.001
		Lecture Test 2	394	0.180	1.161	0.059	0.645	0.123	0.552	0.245	0.101	394	<<0.001	0.961	394	<<0.001
		Exam Test	486	0.136	1.239	0.056	0.165	0.111	0.335	0.221	0.119	486	<<0.001	0.970	486	<<0.001
		Redeemable Exam	456	0.599	1.075	0.050	0.653	0.114	1.543	0.228	0.105	456	<<0.001	0.962	456	<<0.001
	2015	Lecture Test 1	429	0.540	1.143	0.055	0.620	0.118	0.211	0.235	0.128	429	<<0.001	0.957	429	<<0.001
		Lecture Test 2	392	0.362	1.124	0.057	0.725	0.123	0.997	0.246	0.119	392	<<0.001	0.952	392	<<0.001
		Exam Test	487	-0.016	1.171	0.053	-0.005	0.111	0.097	0.221	0.093	487	<<0.001	0.976	487	<<0.001
		Redeemable Exam	472	0.676	1.151	0.053	0.853	0.112	0.570	0.224	0.106	472	<<0.001	0.949	472	<<0.001

Table 68: Tests of Normality on Rasch Item Difficulty Measures from Chemistry IB MCQ Assessment Tasks

			Rasch Measures				Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean	Standard Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Chemistry IB Items	2012	Lecture Test 1	15	0.000	0.618	0.160	-0.411	0.580	0.085	1.121	0.175	15	0.200	0.947	15	0.472
		Lecture Test 2	15	0.000	0.703	0.182	-0.087	0.580	-0.861	1.121	0.167	15	0.200	0.951	15	0.548
		Exam Test	9	-0.001	1.025	0.342	-0.028	0.717	-0.651	1.400	0.129	9	0.200	0.985	9	0.986
		Redeemable Exam	30	0.000	0.748	0.137	-0.404	0.427	-0.540	0.833	0.096	30	0.200	0.966	30	0.436
		Stacked	69	0.000	0.751	0.090	-0.084	0.289	-0.110	0.570	0.070	69	0.200	0.991	69	0.893
	2013	Lecture Test 1	15	0.000	0.607	0.157	0.229	0.580	0.233	1.121	0.123	15	0.200	0.962	15	0.732
		Lecture Test 2	15	-0.001	0.778	0.201	-0.318	0.580	-0.717	1.121	0.121	15	0.200	0.964	15	0.757
		Exam Test	10	0.000	1.127	0.356	0.367	0.687	-1.385	1.334	0.205	10	0.200	0.911	10	0.289
		Redeemable Exam	30	0.000	0.704	0.129	0.152	0.427	-0.350	0.833	0.095	30	0.200	0.979	30	0.802
		Stacked	70	0.000	0.760	0.091	0.335	0.287	-0.362	0.566	0.069	70	0.200	0.984	70	0.495
	2014	Lecture Test 1	15	0.000	0.640	0.165	-0.702	0.580	1.290	1.121	0.183	15	0.188	0.934	15	0.313
		Lecture Test 2	15	0.000	0.769	0.199	-0.197	0.580	-0.877	1.121	0.138	15	0.200	0.956	15	0.628
		Exam Test	10	0.001	1.033	0.327	0.333	0.687	-1.354	1.334	0.137	10	0.200	0.931	10	0.461
		Redeemable Exam	30	0.000	0.673	0.123	-0.245	0.427	0.280	0.833	0.135	30	0.173	0.972	30	0.591
		Stacked	70	0.000	0.726	0.087	0.134	0.287	0.018	0.566	0.101	70	0.074	0.987	70	0.670
	2015	Lecture Test 1	15	0.001	0.668	0.172	-0.716	0.580	1.221	1.121	0.172	15	0.200	0.940	15	0.380
		Lecture Test 2	15	0.001	0.985	0.254	-0.464	0.580	0.702	1.121	0.135	15	0.200	0.978	15	0.954
		Exam Test	10	0.000	0.898	0.284	0.416	0.687	-1.255	1.334	0.168	10	0.200	0.915	10	0.321
		Redeemable Exam	30	0.000	0.770	0.141	-0.666	0.427	1.616	0.833	0.097	30	0.200	0.960	30	0.308
		Stacked	70	0.000	0.795	0.095	-0.260	0.287	0.886	0.566	0.077	70	0.200	0.981	70	0.372

Table 69: Tests of Normality on Rasch Student Ability Measures from Foundations of Chemistry IA MCQ Assessment Tasks

			Rasch Measures				Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean	Standard Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Foundations of Chemistry IA Students	**2012**	**Lecture Test 1**	259	1.637	1.209	0.075	0.201	0.151	-0.483	0.302	0.143	259	<<0.001	0.958	259	<<0.001
		Lecture Test 2	267	0.699	1.267	0.078	0.208	0.149	-0.189	0.297	0.092	267	<<0.001	0.980	267	0.001
		Exam Test	306	-0.035	1.219	0.070	0.205	0.139	0.539	0.278	0.131	306	<<0.001	0.969	306	<<0.001
		Redeemable Exam	258	0.779	1.038	0.065	0.205	0.139	0.539	0.278	0.081	258	<<0.001	0.990	258	0.066
	2013	**Lecture Test 1**	309	1.596	1.261	0.072	0.187	0.139	-0.416	0.276	0.112	309	<<0.001	0.965	309	<<0.001
		Lecture Test 2	255	0.453	1.299	0.081	0.187	0.153	0.173	0.304	0.076	255	0.001	0.986	255	0.012
		Exam Test	365	-0.135	1.254	0.066	0.416	0.128	0.330	0.255	0.111	365	<<0.001	0.966	365	<<0.001
		Redeemable Exam	336	0.879	1.056	0.058	0.702	0.133	0.595	0.265	0.099	336	<<0.001	0.964	336	<<0.001
	2014	**Lecture Test 1**	252	1.612	1.358	0.086	-0.081	0.153	-0.300	0.306	0.106	252	<<0.001	0.969	252	<<0.001
		Lecture Test 2	223	0.640	1.203	0.081	0.304	0.163	-0.122	0.324	0.119	223	<<0.001	0.972	223	<<0.001
		Exam Test	327	-0.275	1.295	0.072	0.505	0.135	0.359	0.269	0.128	327	<<0.001	0.961	327	<<0.001
		Redeemable Exam	301	0.802	1.175	0.068	0.531	0.140	1.028	0.280	0.079	301	<<0.001	0.971	301	<<0.001
	2015	**Lecture Test 1**	294	1.605	1.389	0.081	0.089	0.142	-0.587	0.283	0.111	294	<<0.001	0.966	294	<<0.001
		Lecture Test 2	236	0.655	1.337	0.087	-0.044	0.158	0.302	0.316	0.093	236	<<0.001	0.984	236	0.011
		Exam Test	367	-0.073	1.266	0.066	0.341	0.127	0.433	0.254	0.108	367	<<0.001	0.969	367	<<0.001
		Redeemable Exam	331	0.736	1.140	0.063	0.303	0.134	0.688	0.267	0.096	331	<<0.001	0.986	331	0.002

Table 70: Tests of Normality on Rasch Item Difficulty Measures from Foundations of Chemistry IA MCQ Assessment Tasks

			Rasch Measures				Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean	Standard Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Foundations of Chemistry IA Items	2012	Lecture Test 1	15	0.001	1.148	0.297	0.260	0.580	-0.092	1.121	0.139	15	0.200	0.947	15	0.473
		Lecture Test 2	15	0.000	1.503	0.388	0.430	0.568	-0.103	1.121	0.153	15	0.200	0.959	15	0.674
		Exam Test	10	0.000	1.068	0.338	0.389	0.687	-0.681	1.334	0.193	10	0.200	0.942	10	0.581
		Redeemable Exam	30	-0.002	1.405	0.256	-0.552	0.427	-0.509	0.833	0.153	30	0.071	0.938	30	0.081
		Stacked	70	0.000	1.323	0.158	-0.122	0.287	-0.351	0.566	0.095	70	0.188	0.976	70	0.195
	2013	Lecture Test 1	15	0.000	1.277	0.330	0.875	0.580	0.107	1.121	0.191	15	0.147	0.904	15	0.109
		Lecture Test 2	15	-0.001	1.510	0.390	0.329	0.580	0.216	1.121	0.187	15	0.164	0.943	15	0.421
		Exam Test	10	0.001	1.149	0.363	0.709	0.687	0.074	1.334	0.142	10	0.200	0.951	10	0.679
		Redeemable Exam	30	0.000	1.459	0.266	-0.318	0.427	-0.144	0.833	0.110	30	0.200	0.969	30	0.524
		Stacked	70	-0.001	1.324	0.158	0.168	0.287	-0.325	0.566	0.065	70	0.200	0.975	70	0.184
	2014	Lecture Test 1	15	0.001	1.217	0.314	0.974	0.580	0.891	1.121	0.141	15	0.200	0.921	15	0.202
		Lecture Test 2	15	0.000	1.548	0.400	0.241	0.580	0.001	1.121	0.193	15	0.137	0.962	15	0.722
		Exam Test	10	0.001	1.419	0.400	-0.335	0.687	1.360	1.334	0.229	10	0.147	0.890	10	0.169
		Redeemable Exam	30	0.000	1.419	0.259	-0.335	0.427	-0.409	0.833	0.127	30	0.200	0.960	30	0.312
		Stacked	70	0.000	1.385	0.166	0.135	0.287	-0.064	0.566	0.087	70	0.200	0.985	70	0.582
	2015	Lecture Test 1	15	0.000	1.258	0.325	0.858	0.580	0.179	1.121	0.147	15	0.200	0.908	15	0.125
		Lecture Test 2	15	-0.001	1.499	0.387	0.207	0.580	-0.341	1.121	0.127	15	0.200	0.971	15	0.877
		Exam Test	10	0.000	1.210	0.383	0.655	0.687	0.032	1.334	0.223	10	0.172	0.938	10	0.533
		Redeemable Exam	30	-0.001	1.361	0.249	-0.212	0.427	-0.259	0.833	0.100	30	0.200	0.971	30	0.575
		Stacked	70	0.001	1.319	0.158	0.126	0.287	-0.369	0.566	0.075	70	0.200	0.984	70	0.493

Table 71: Tests of Normality on Rasch Student Ability Measures from Foundations of Chemistry IB MCQ Assessment Tasks

			Rasch Measures				Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean	Standard Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Foundations of Chemistry IB Students	2012	Lecture Test 1	236	1.188	1.480	0.096	0.239	0.158	-0.434	0.316	0.093	236	<<0.001	0.972	236	<<0.001
		Lecture Test 2	189	0.075	1.292	0.094	0.098	0.177	-0.040	0.352	0.095	189	<<0.001	0.980	189	0.010
		Exam Test	266	0.919	1.241	0.076	0.219	0.149	0.039	0.298	0.146	266	<<0.001	0.962	266	<<0.001
		Redeemable Exam	250	0.839	1.352	0.085	0.708	0.154	0.908	0.307	0.127	250	<<0.001	0.959	250	<<0.001
	2013	Lecture Test 1	249	0.864	1.321	0.084	0.273	0.154	-0.323	0.307	0.095	249	<<0.001	0.975	249	<<0.001
		Lecture Test 2	218	0.527	1.243	0.084	0.132	0.165	-0.553	0.328	0.101	218	<<0.001	0.979	218	0.002
		Exam Test	305	-0.539	1.262	0.072	0.177	0.140	0.220	0.278	0.115	305	<<0.001	0.969	305	<<0.001
		Redeemable Exam	288	1.067	1.313	0.077	0.517	0.144	0.012	0.286	0.063	288	0.007	0.972	288	<<0.001
	2014	Lecture Test 1	216	0.874	1.391	0.095	0.030	0.166	-0.137	0.330	0.085	216	0.001	0.983	216	0.010
		Lecture Test 2	184	0.938	1.446	0.107	0.315	0.179	-0.106	0.356	0.136	184	<<0.001	0.970	184	<<0.001
		Exam Test	276	0.066	1.219	0.073	0.505	0.147	0.466	0.292	0.145	276	<<0.001	0.957	276	<<0.001
		Redeemable Exam	259	0.591	1.196	0.074	0.635	0.151	0.381	0.302	0.088	259	<<0.001	0.971	259	<<0.001
	2015	Lecture Test 1	231	0.667	1.298	0.085	0.195	0.160	0.041	0.319	0.111	231	<<0.001	0.978	231	0.001
		Lecture Test 2	198	0.842	1.336	0.095	0.359	0.173	-0.223	0.344	0.114	198	<<0.001	0.974	198	0.001
		Exam Test	300	0.163	1.184	0.068	0.266	0.141	0.137	0.281	0.102	300	<<0.001	0.969	300	<<0.001
		Redeemable Exam	277	0.662	1.144	0.069	0.271	0.146	0.199	0.292	0.061	277	0.015	0.987	277	0.017

Table 72: Tests of Normality on Rasch Item Difficulty Measures from Foundations of Chemistry IB MCQ Assessment Tasks

			Rasch Measures				Skewness		Kurtosis		Kolmogorov-Smirnov			Shapiro-Wilk		
			n	Mean	Standard Deviation	Std. Error	Value	Std. Error	Value	Std. Error	Statistic	d.f.	*p*-value	Statistic	d.f.	*p*-value
Foundations of Chemistry IB Items	2012	Lecture Test 1	15	0.001	1.320	0.341	0.880	0.580	0.476	1.121	0.170	15	0.200	0.931	15	0.281
		Lecture Test 2	14	-0.001	1.304	0.349	0.804	0.597	-0.184	1.154	0.179	14	0.200	0.916	14	0.193
		Exam Test	10	0.000	1.562	0.494	0.323	0.687	-1.357	1.334	0.183	10	0.200	0.916	10	0.322
		Redeemable Exam	30	0.000	1.289	0.235	-0.075	0.427	0.326	0.833	0.124	30	0.200	0.976	30	0.716
		Stacked	69	0.000	1.294	0.156	0.280	0.289	-0.041	0.570	0.066	69	0.200	0.987	69	0.677
	2013	Lecture Test 1	15	0.001	1.340	0.346	0.067	0.580	-0.778	1.121	0.160	15	0.200	0.956	15	0.618
		Lecture Test 2	15	0.001	1.114	0.288	0.453	0.580	-0.316	1.121	0.124	15	0.200	0.966	15	0.789
		Exam Test	10	0.001	1.033	0.327	-0.385	0.687	0.460	1.334	0.118	10	0.200	0.979	10	0.962
		Redeemable Exam	30	0.000	1.251	0.236	-0.098	0.441	0.553	0.858	0.101	28	0.200	0.985	28	0.949
		Stacked	68	0.000	1.453	0.176	0.644	0.291	1.413	0.574	0.070	68	0.200	0.967	68	0.071
	2014	Lecture Test 1	15	0.001	1.223	0.316	0.433	0.580	-0.046	1.121	0.146	15	0.200	0.946	15	0.470
		Lecture Test 2	15	0.000	1.484	0.383	-0.170	0.580	-0.148	1.121	0.137	15	0.200	0.966	15	0.798
		Exam Test	10	0.000	1.222	0.387	-0.045	0.687	-1.436	1.334	0.165	10	0.200	0.915	10	0.317
		Redeemable Exam	30	0.001	1.325	0.242	-0.548	0.427	0.767	0.833	0.078	30	0.200	0.957	30	0.263
		Stacked	70	0.000	1.239	0.148	-0.270	0.287	0.138	0.566	0.078	70	0.200	0.977	70	0.230
	2015	Lecture Test 1	15	0.001	1.219	0.315	0.234	0.580	-0.474	1.121	0.132	15	0.200	0.968	15	0.822
		Lecture Test 2	15	0.001	1.420	0.367	-0.086	0.580	-0.146	1.121	0.156	15	0.200	0.970	15	0.860
		Exam Test	10	-0.001	1.197	0.379	-0.094	0.687	-1.110	1.334	0.153	10	0.200	0.939	10	0.544
		Redeemable Exam	30	0.000	1.170	0.214	-0.303	0.427	0.704	0.833	0.089	30	0.200	0.977	30	0.739
		Stacked	70	0.000	1.155	0.138	-0.134	0.287	-0.078	0.566	0.065	70	0.200	0.984	70	0.519

7.7 MCQ Assessment Rasch Student Ability Histogram and Q-Q Plot

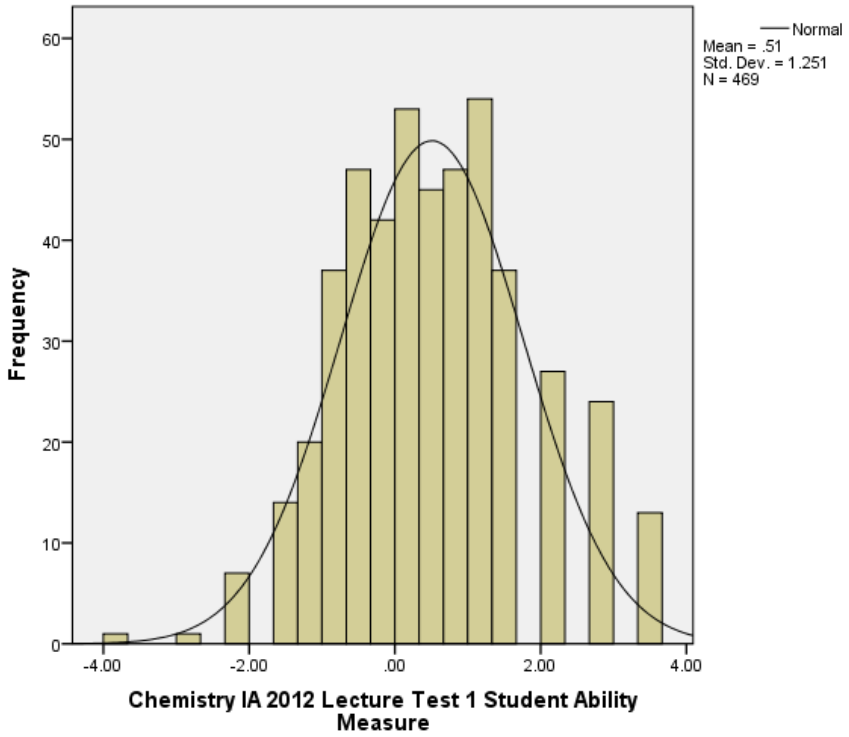

Figure 237: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IA 2012 to Determine the Distribution that the Measures Follow

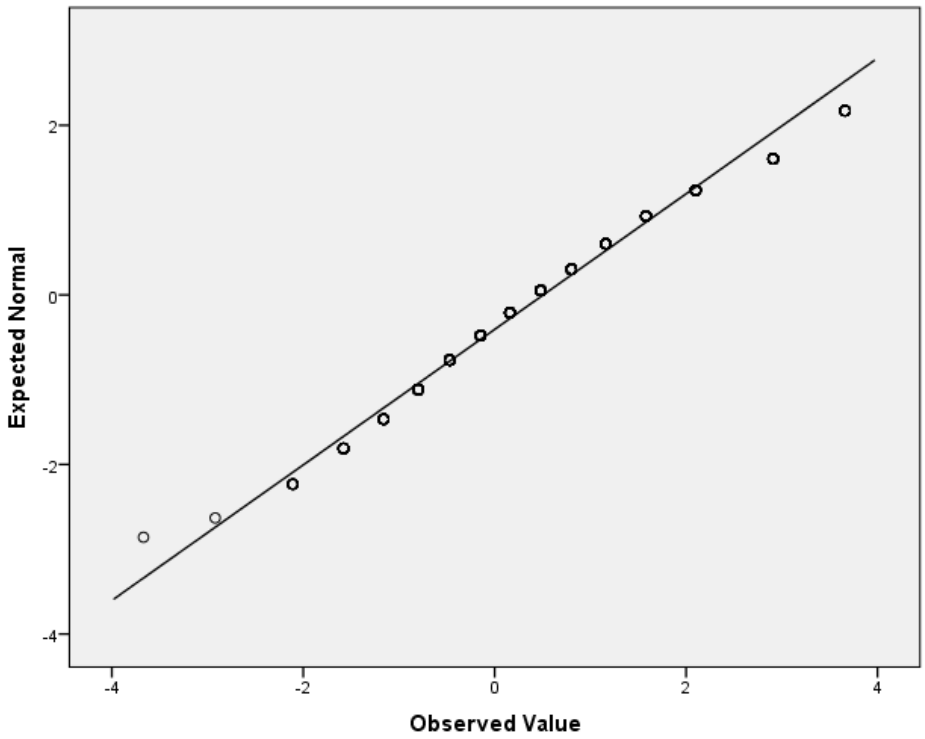

Figure 238: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Lecture Test 1 2012

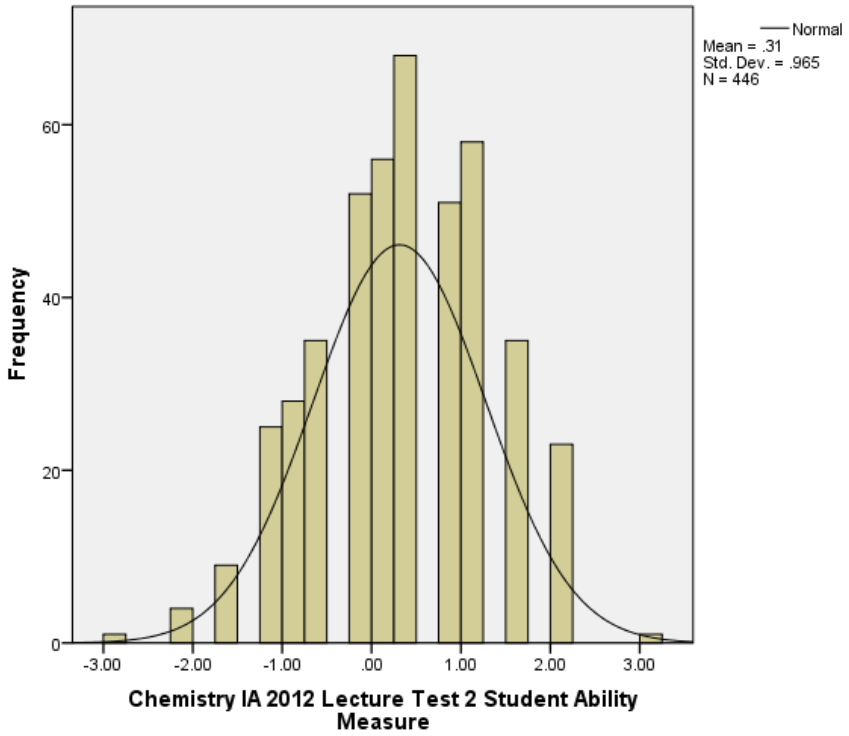

Figure 239: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Chemistry IA 2012 to Determine the Distribution that the Measures Follow

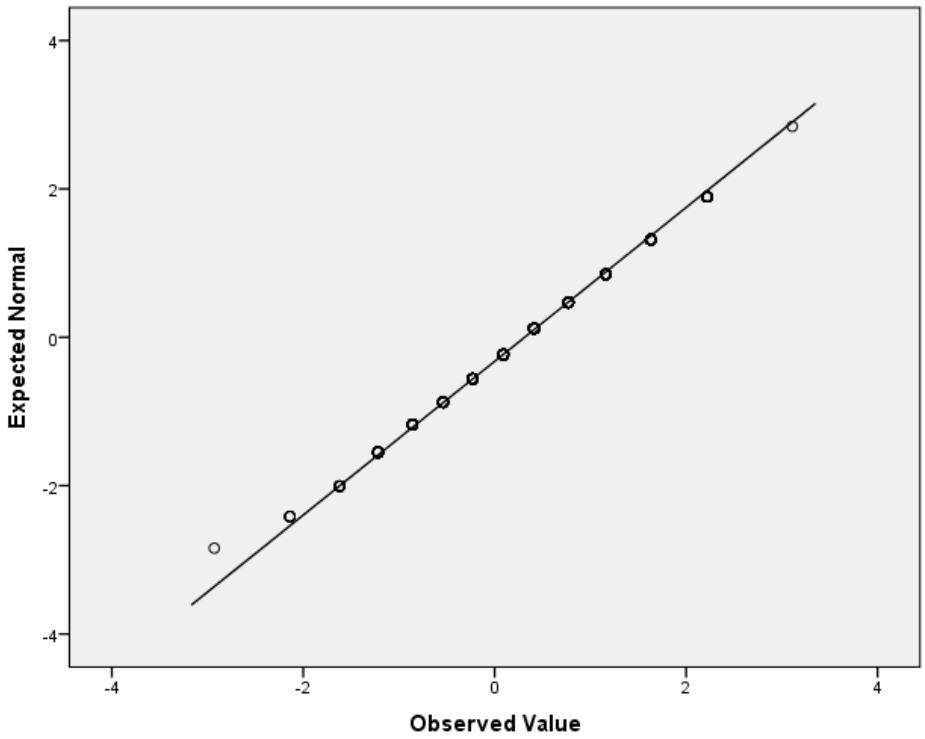

Figure 240: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Lecture Test 2 2012

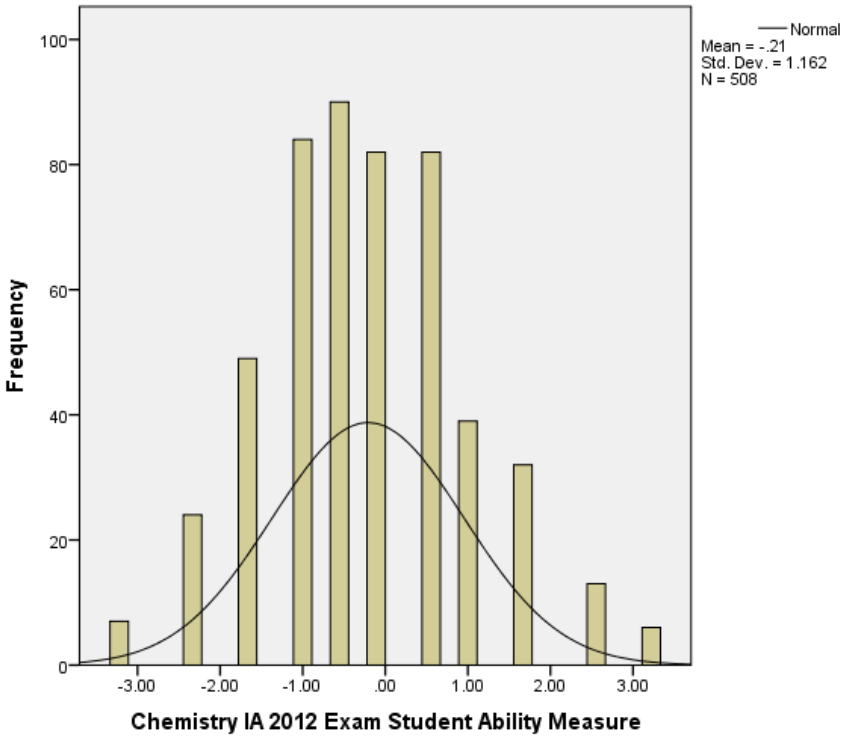

Figure 241: Histogram of the Rasch Student Ability Measures in Exam from Chemistry IA 2012 to Determine the Distribution that the Measures Follow

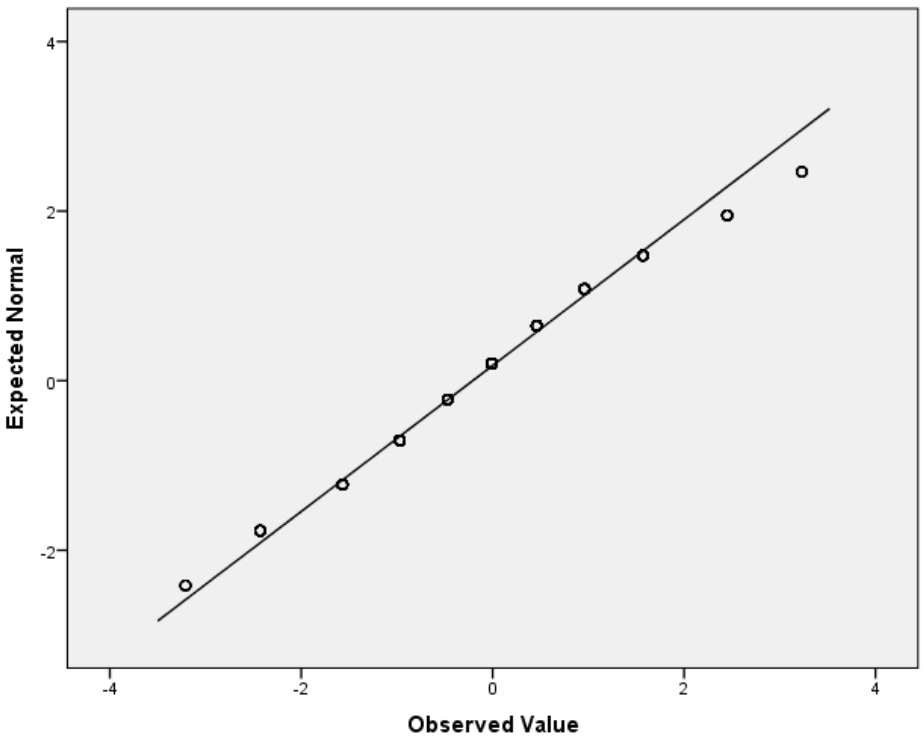

Figure 242: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Exam 2012

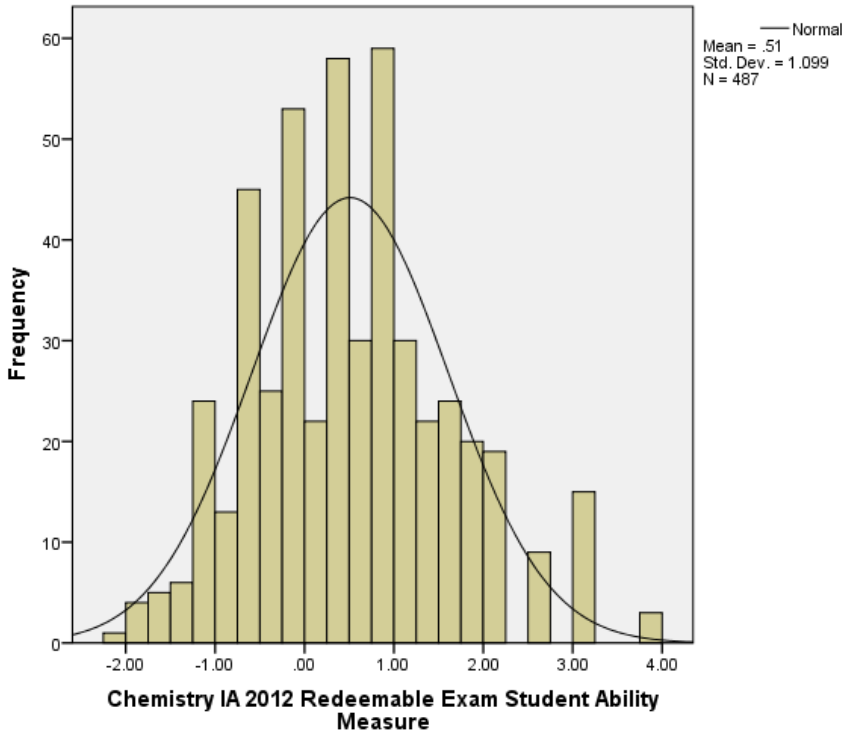

Figure 243: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Chemistry IA 2012 to Determine the Distribution that the Measures Follow

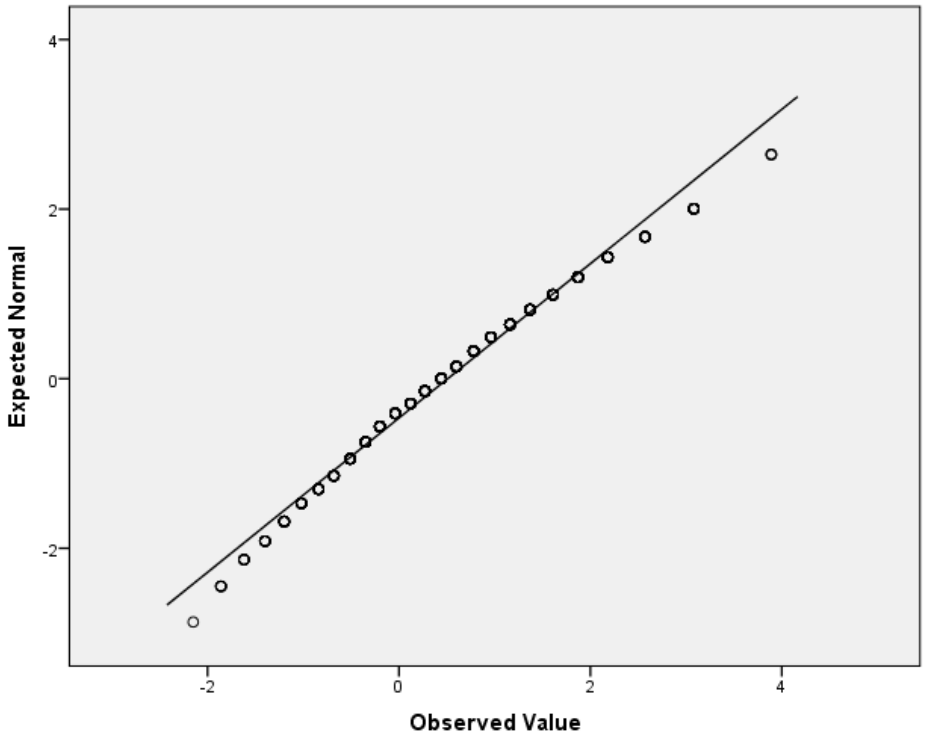

Figure 244: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Redeemable Exam 2012

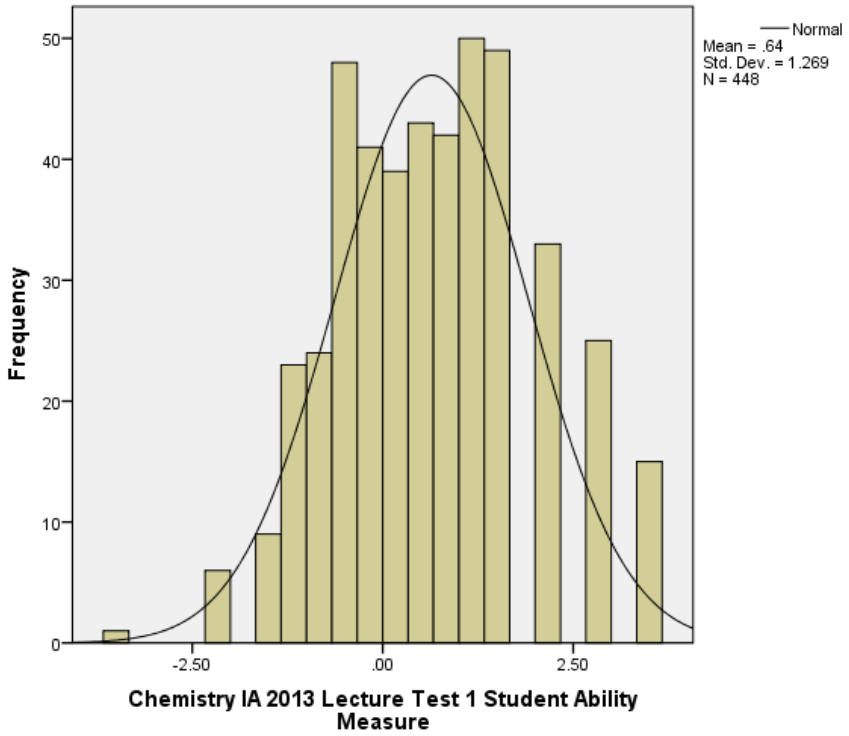

Figure 245: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IA 2013 to Determine the Distribution that the Measures Follow

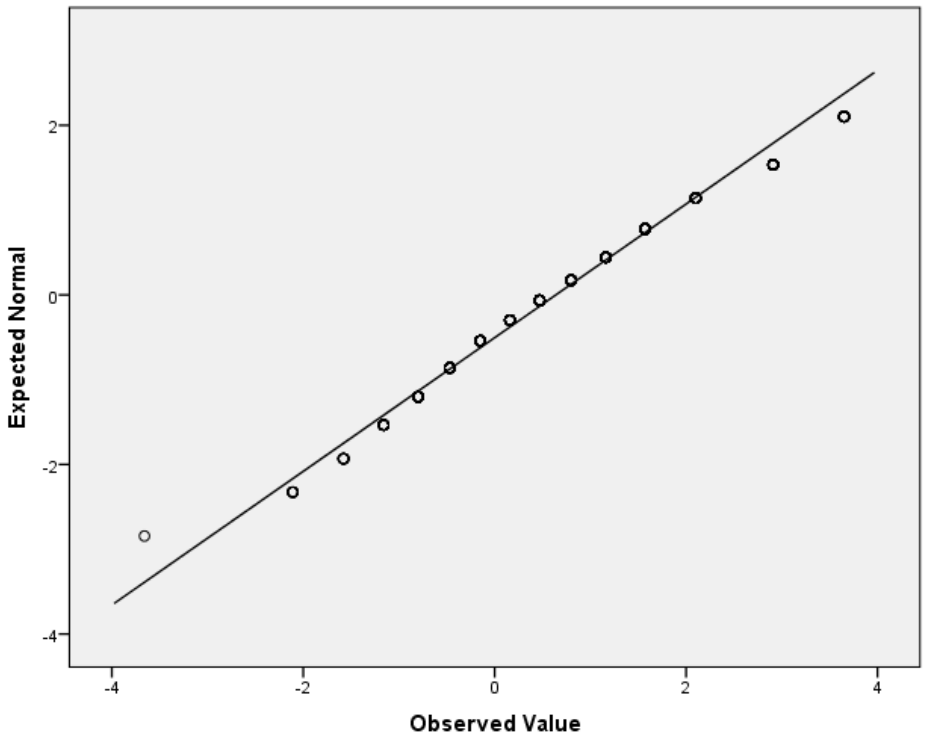

Figure 246: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Lecture Test 1 2013

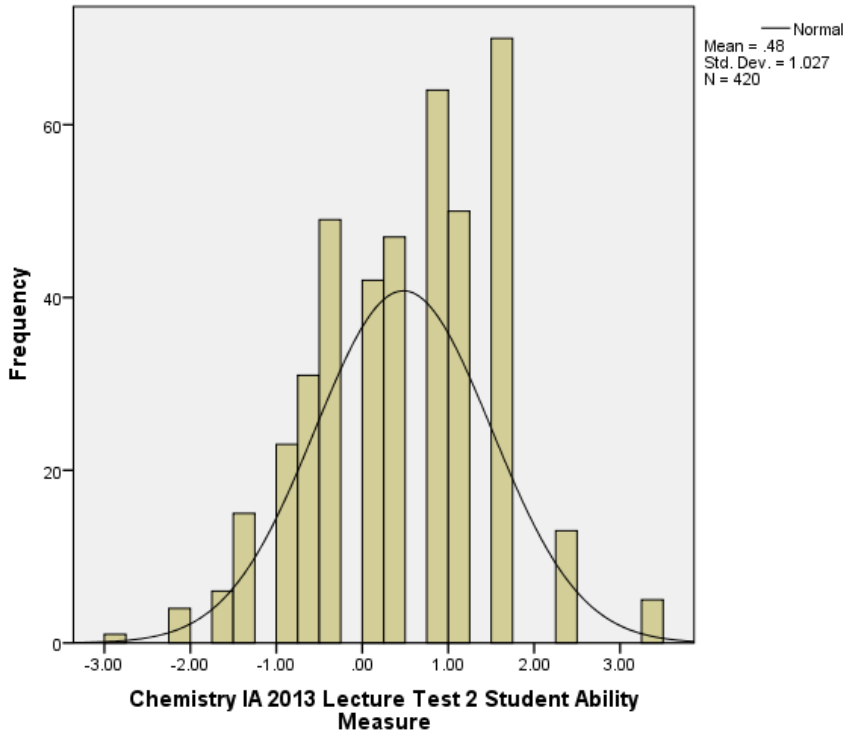

Figure 247: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Chemistry IA 2013 to Determine the Distribution that the Measures Follow

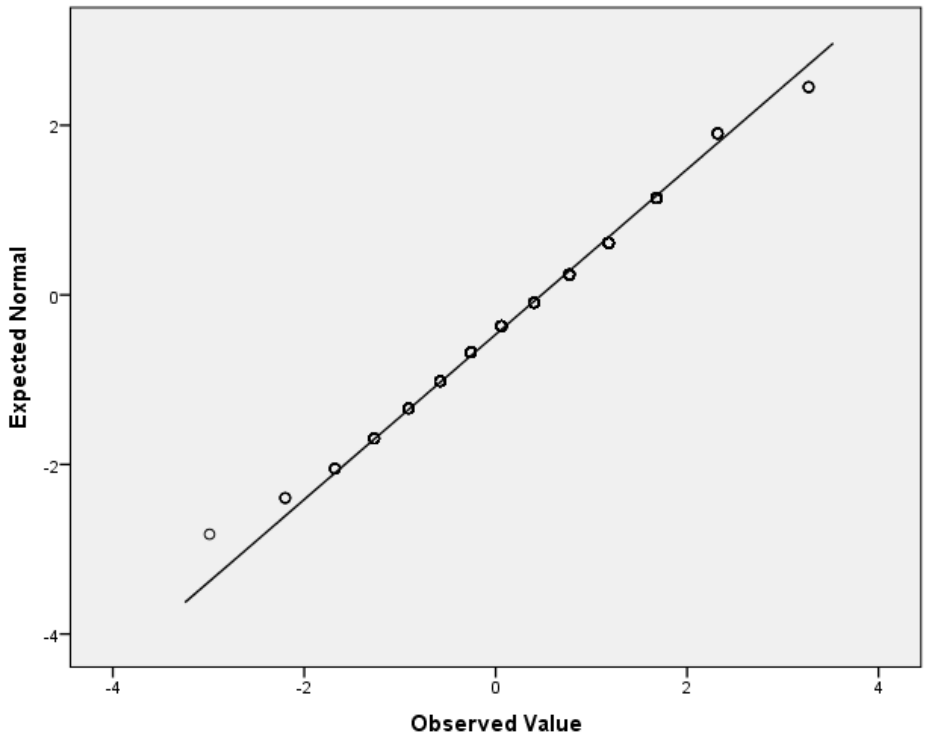

Figure 248: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Lecture Test 2 2013

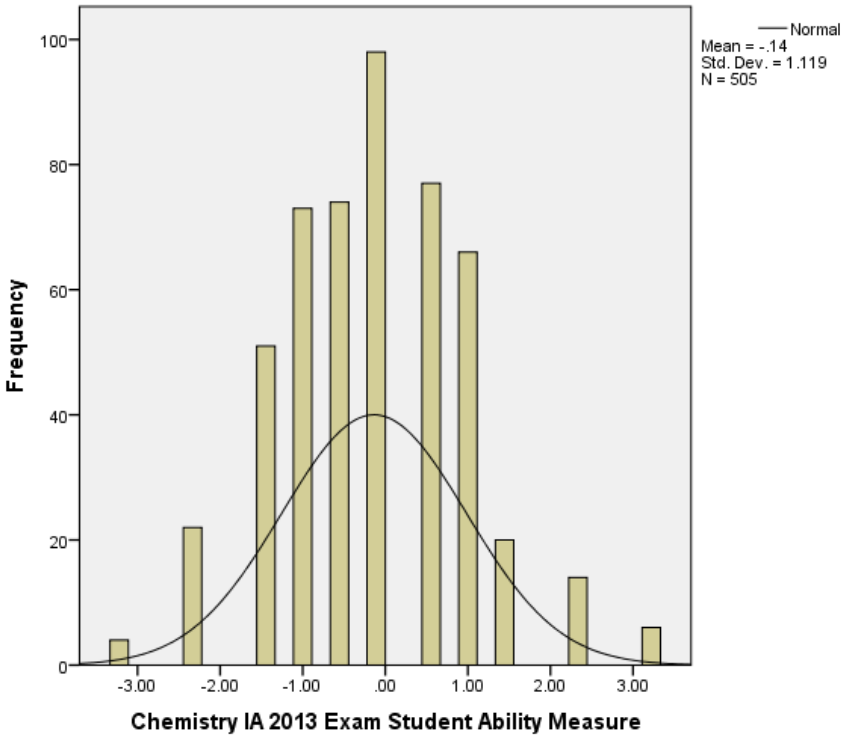

Figure 249: Histogram of the Rasch Student Ability Measures in Exam from Chemistry IA 2013 to Determine the Distribution that the Measures Follow

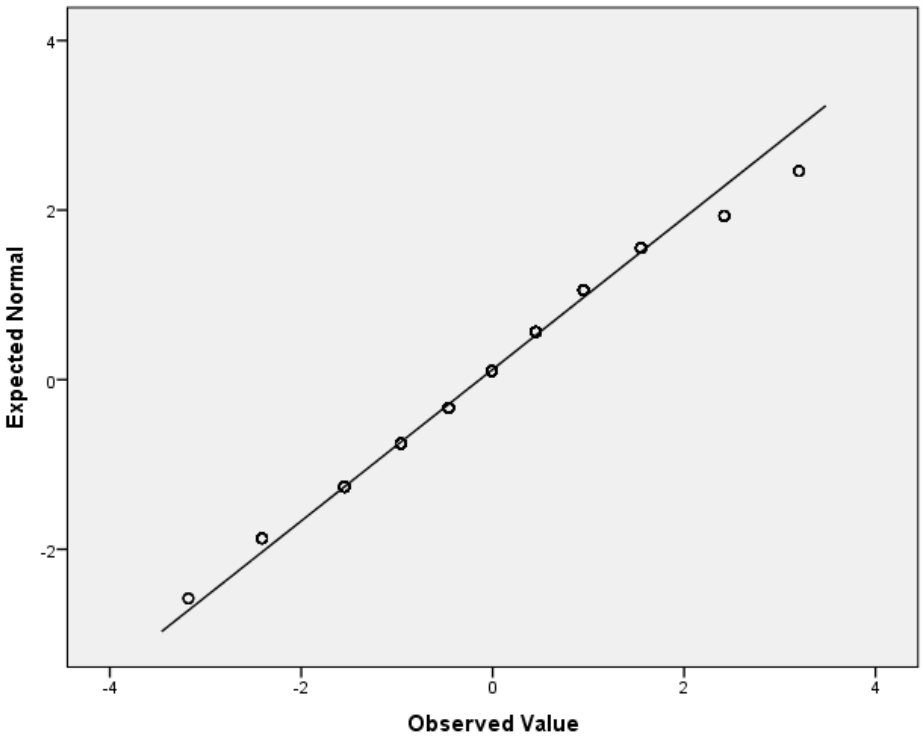

Figure 250: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Exam 2013

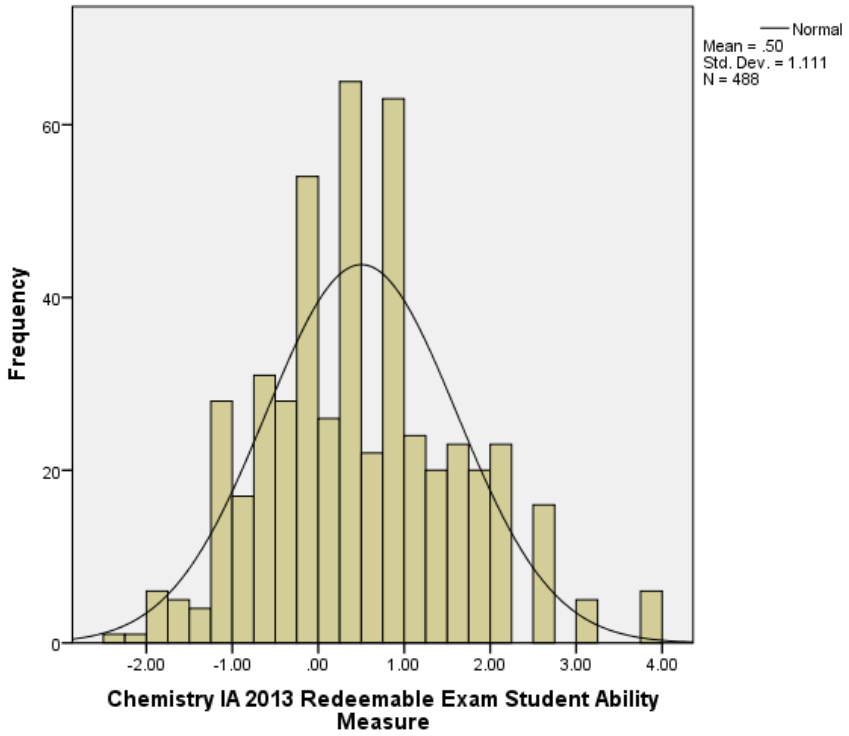

Figure 251: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Chemistry IA 2013 to Determine the Distribution that the Measures Follow

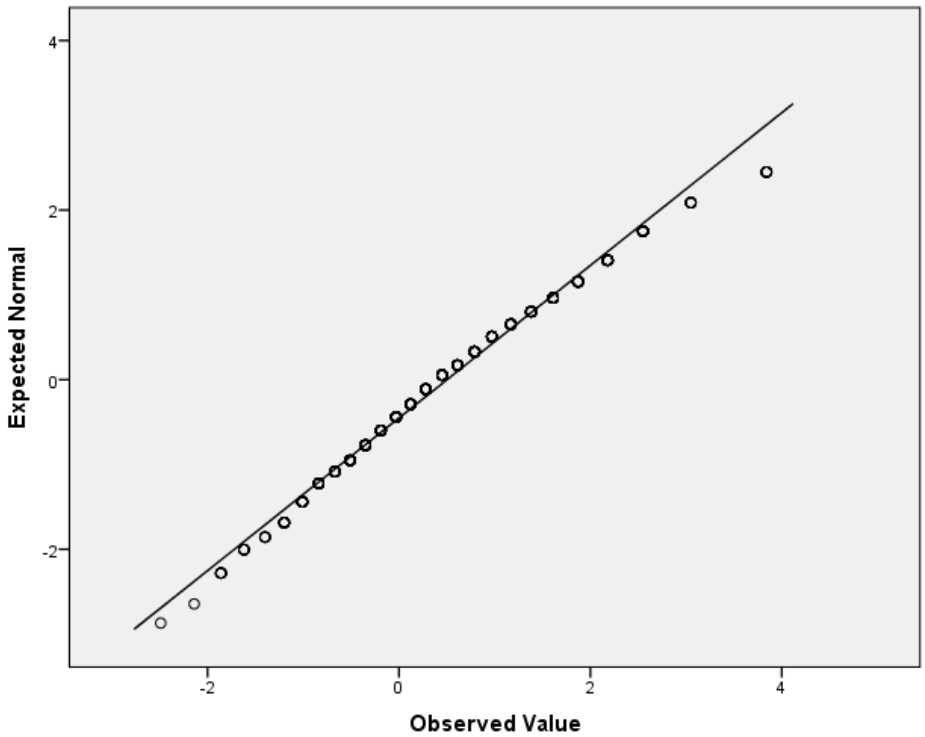

Figure 252: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Redeemable Exam 2013

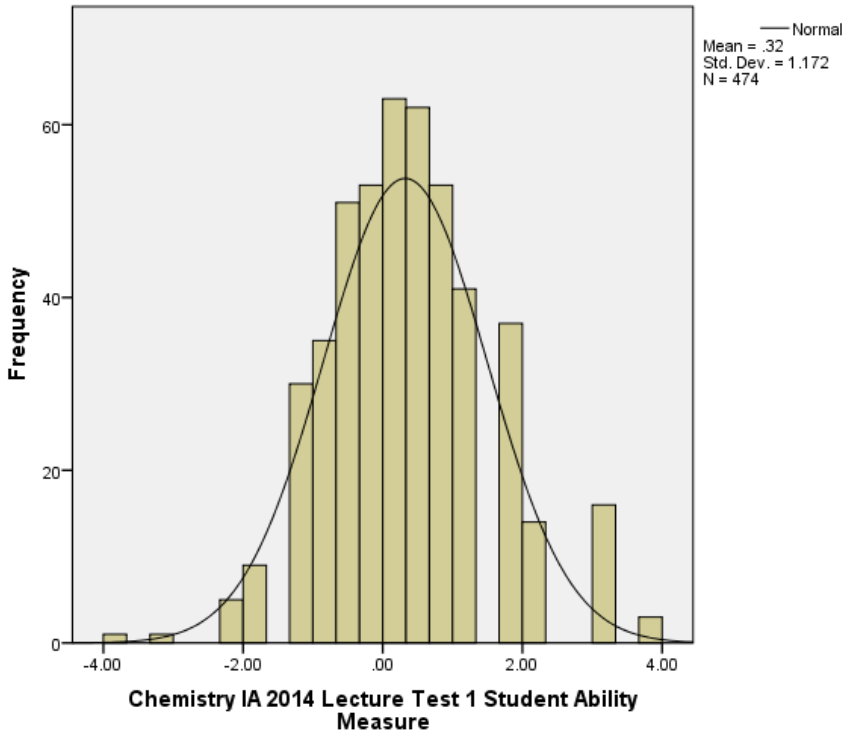

Figure 253: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IA 2014 to Determine the Distribution that the Measures Follow

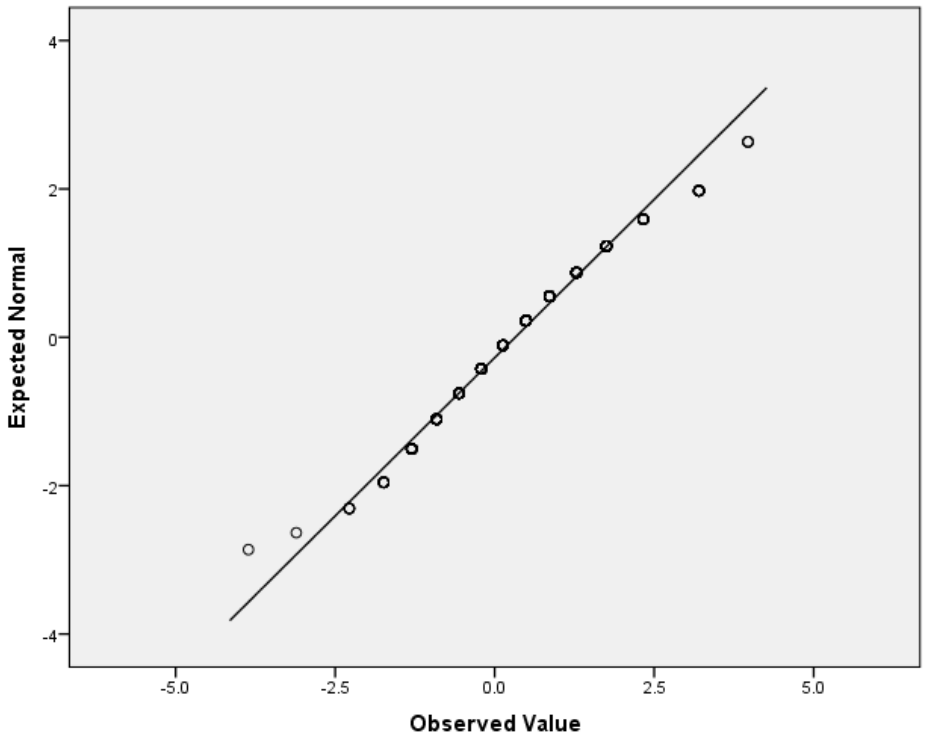

Figure 254: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Lecture Test 1 2014

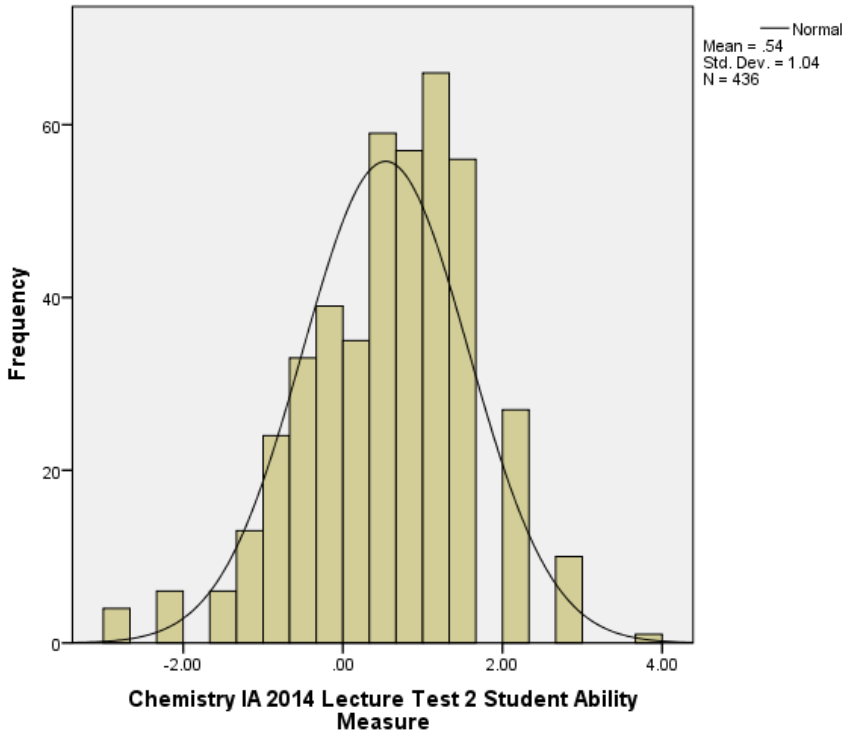

Figure 255: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Chemistry IA 2014 to Determine the Distribution that the Measures Follow

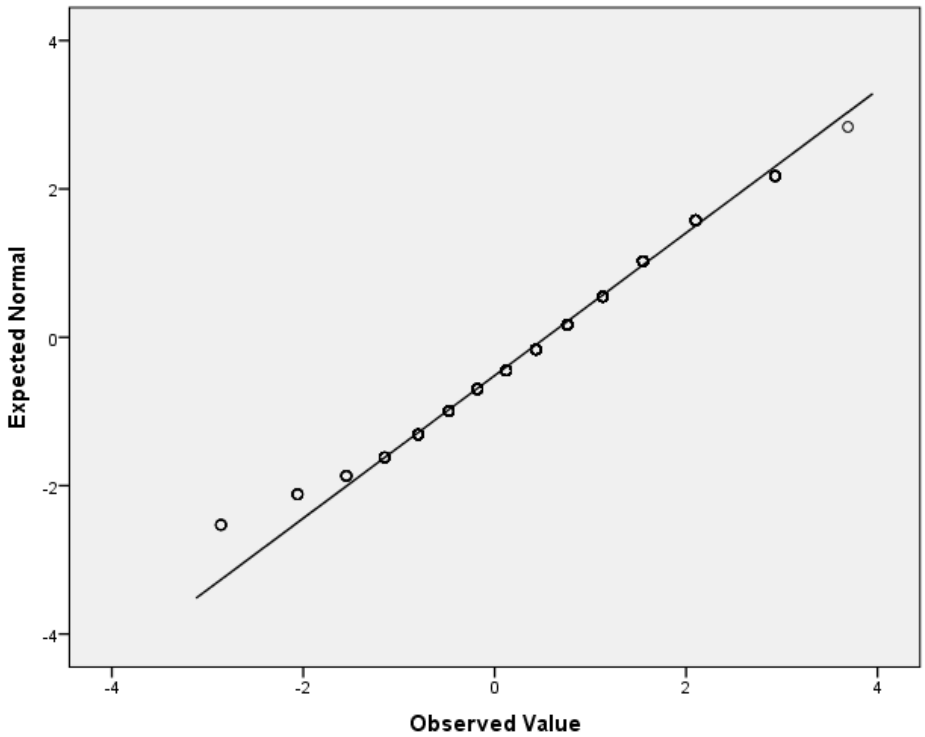

Figure 256: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Lecture Test 2 2014

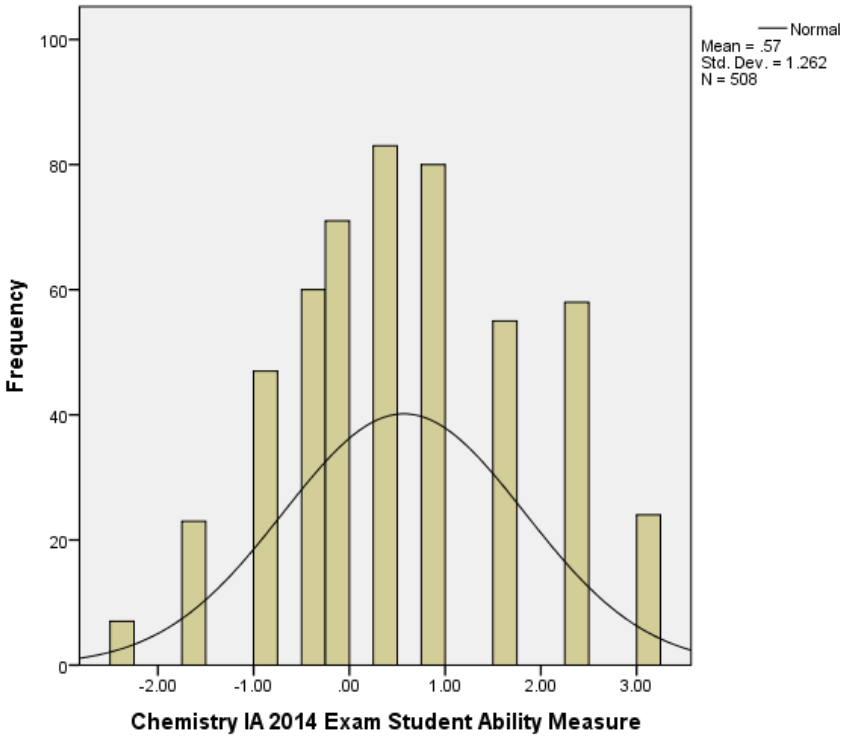

Figure 257: Histogram of the Rasch Student Ability Measures in Exam from Chemistry IA 2014 to Determine the Distribution that the Measures Follow

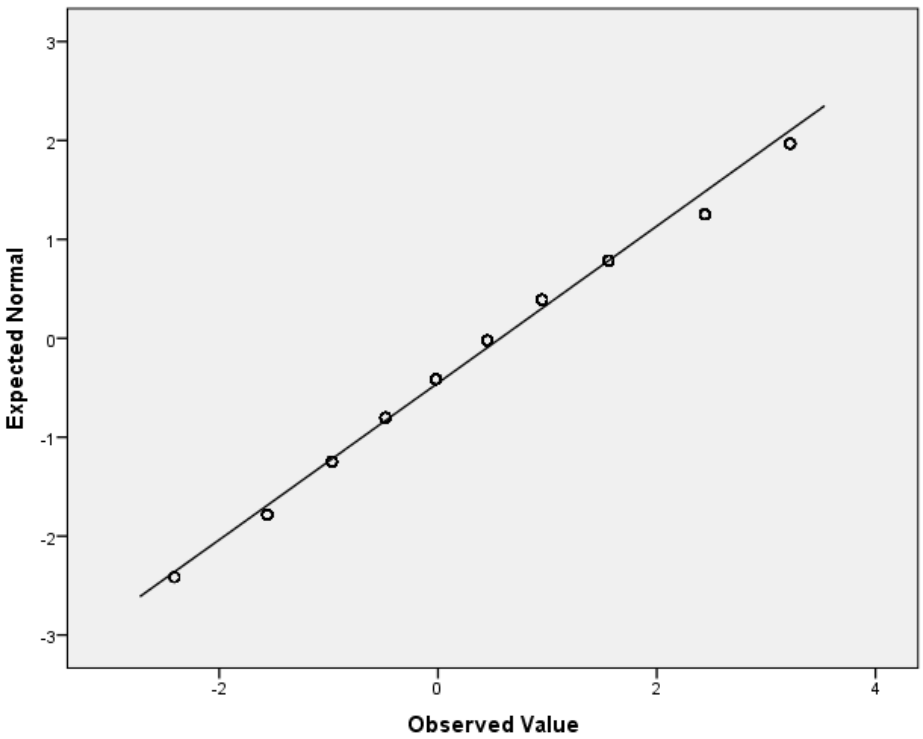

Figure 258: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Exam 2014

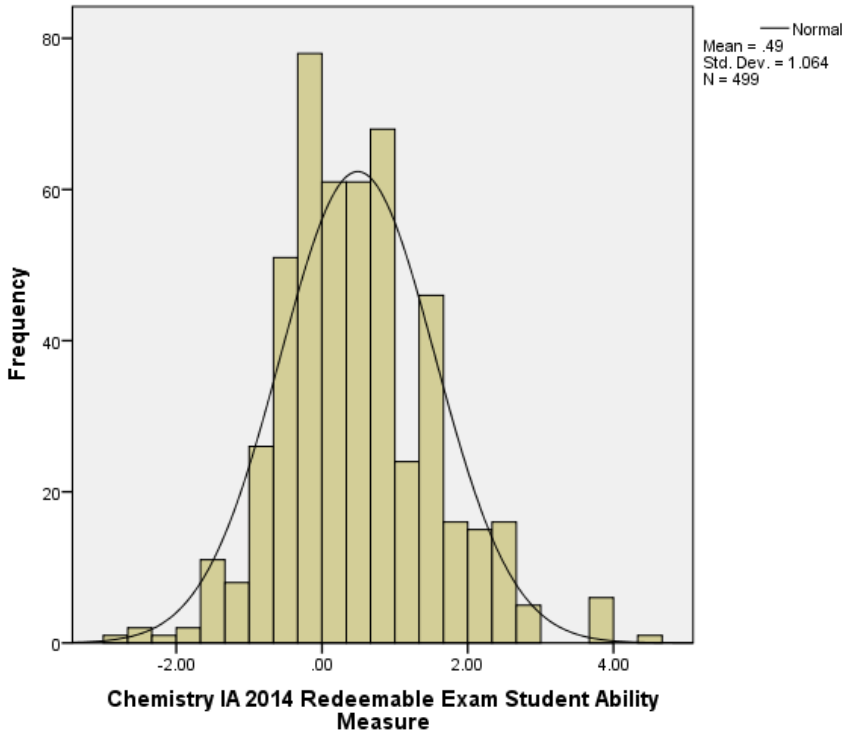

Figure 259: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Chemistry IA 2014 to Determine the Distribution that the Measures Follow

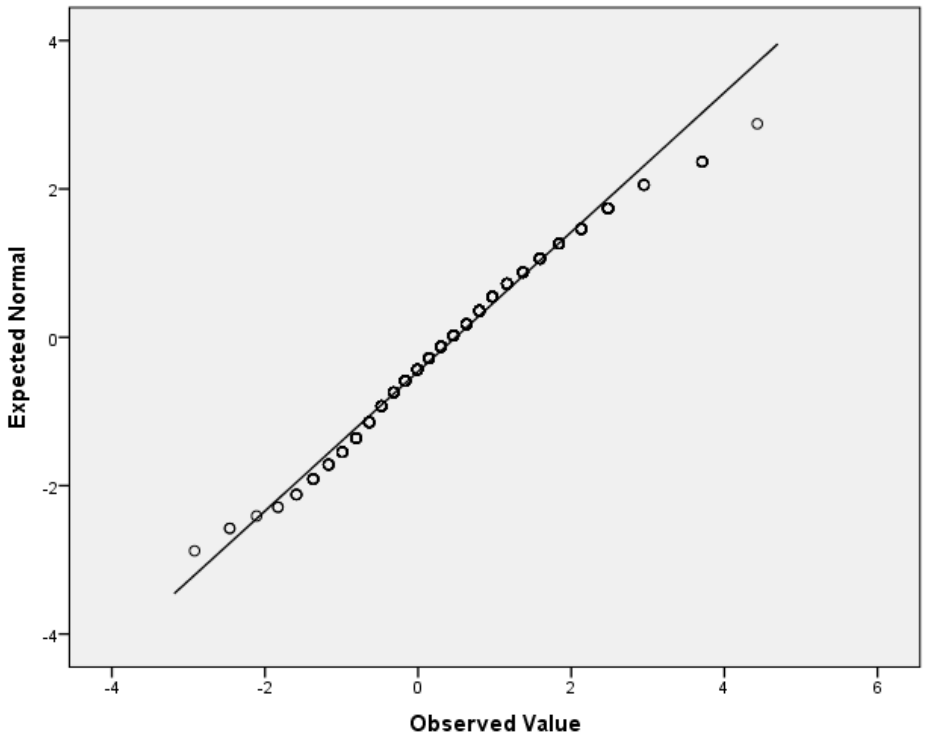

Figure 260: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Redeemable Exam 2014

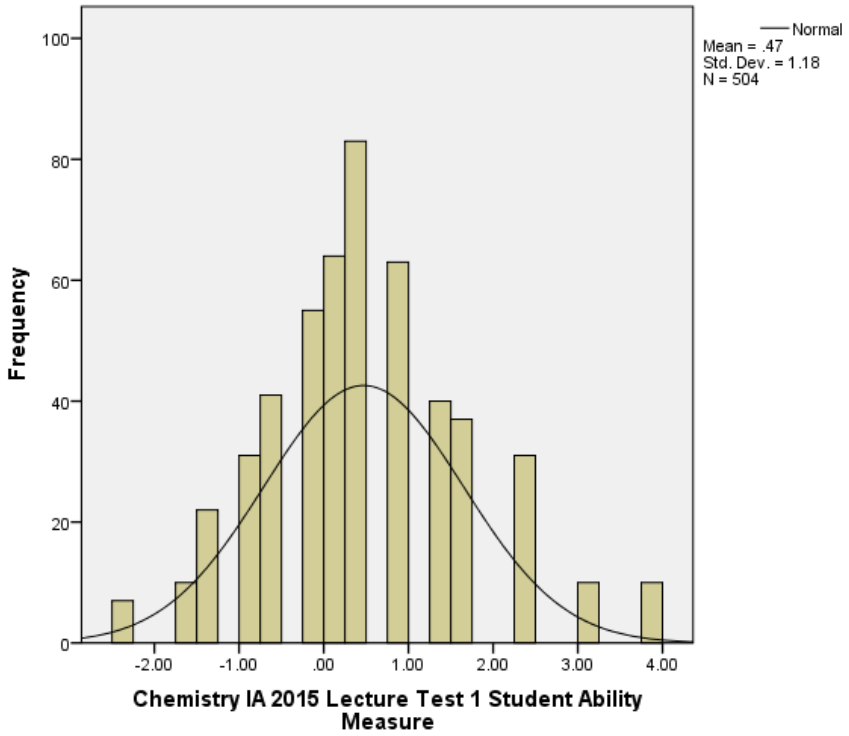

Figure 261: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IA 2015 to Determine the Distribution that the Measures Follow

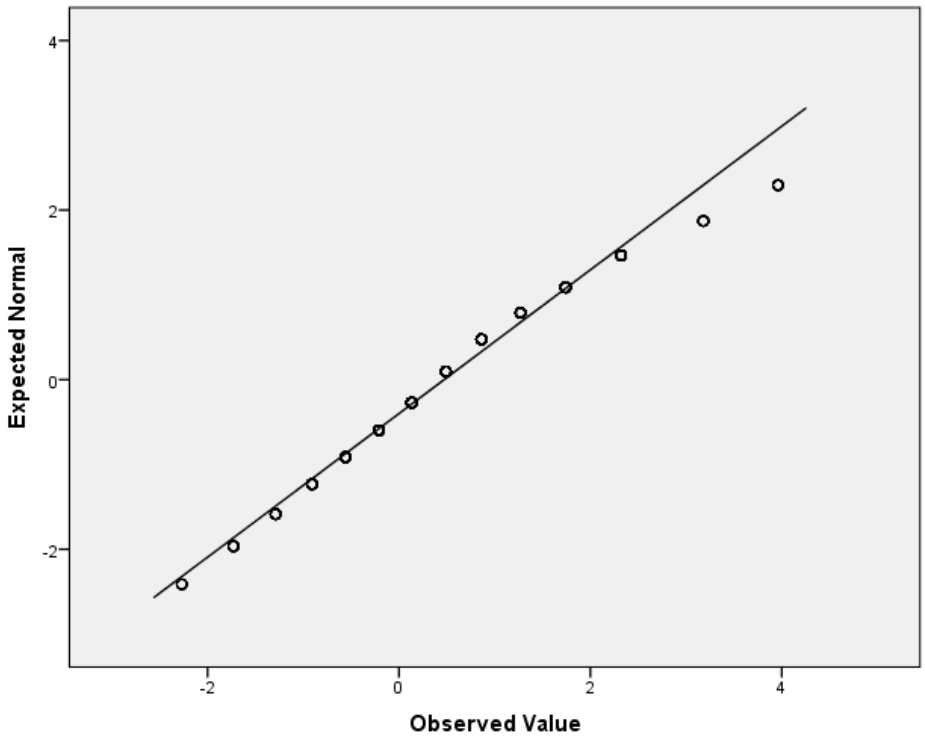

Figure 262: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Lecture Test 1 2015

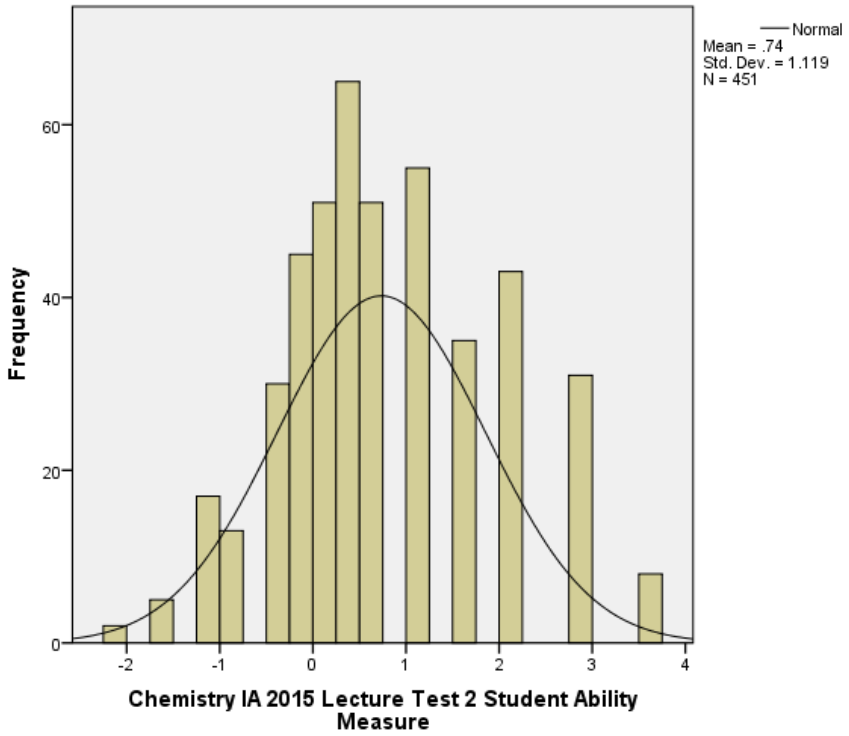

Figure 263: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Chemistry IA 2015 to Determine the Distribution that the Measures Follow

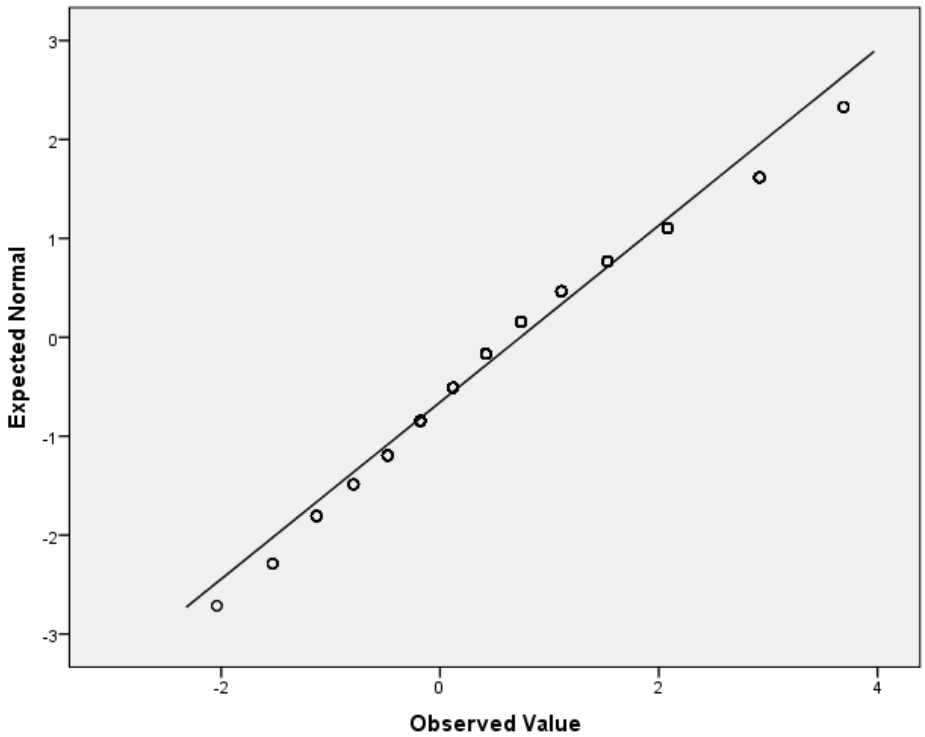

Figure 264: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Lecture Test 2 2015

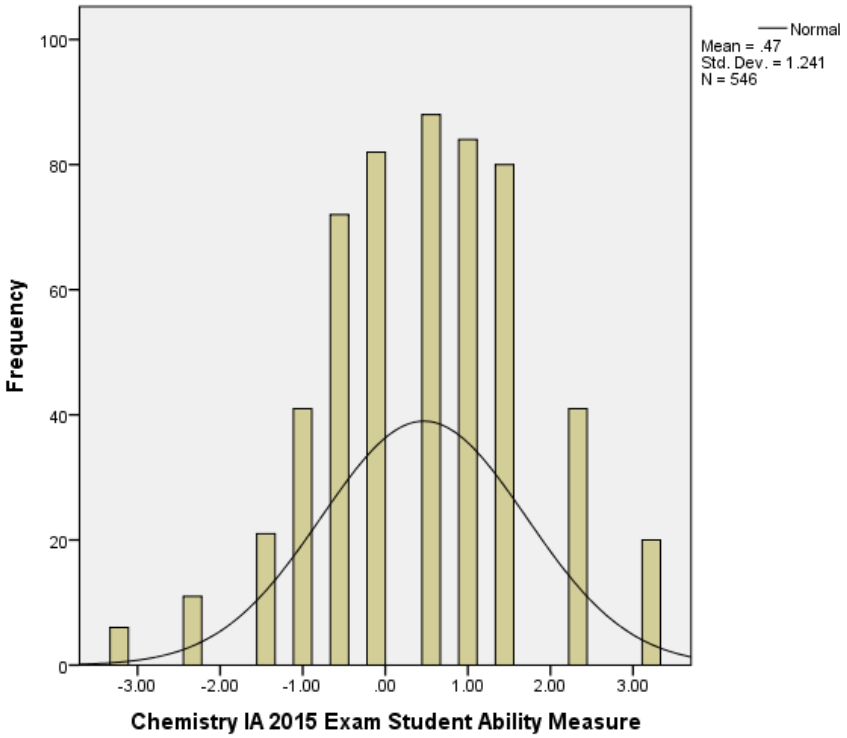

Figure 265: Histogram of the Rasch Student Ability Measures in Exam from Chemistry IA 2015 to Determine the Distribution that the Measures Follow

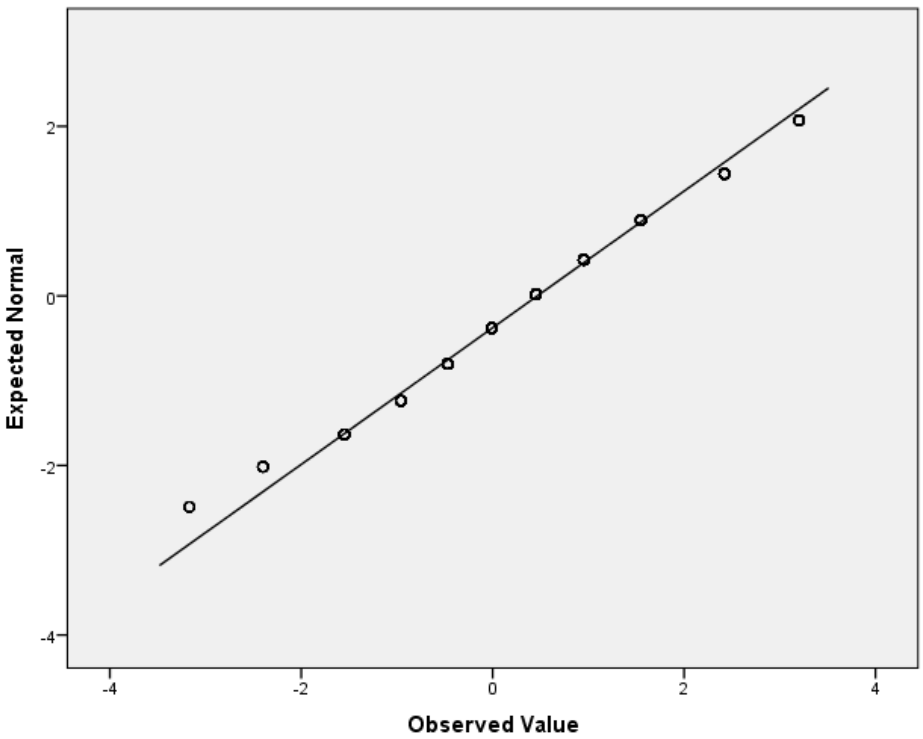

Figure 266: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Exam 2015

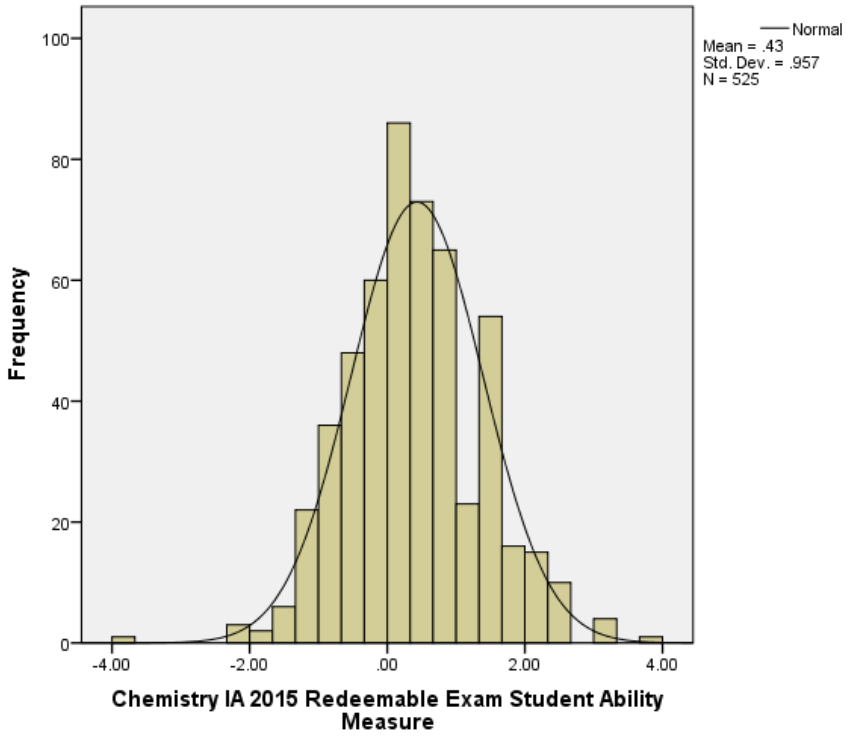

Figure 267: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Chemistry IA 2015 to Determine the Distribution that the Measures Follow

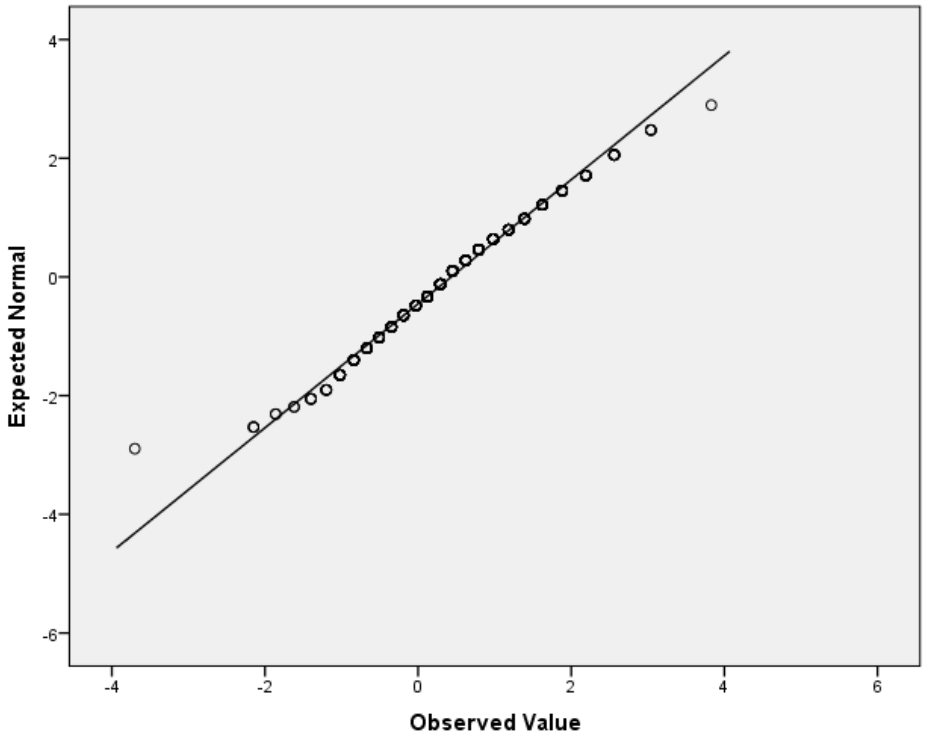

Figure 268: Rasch Student Ability Measure Q-Q Plot from Chemistry IA Redeemable Exam 2015

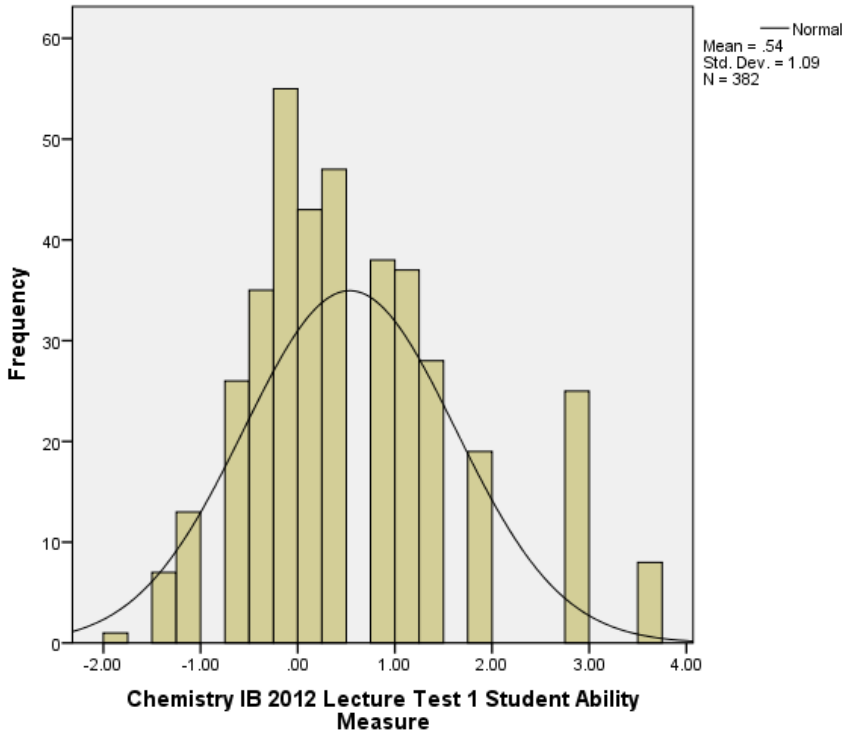

Figure 269: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IB 2012 to Determine the Distribution that the Measures Follow

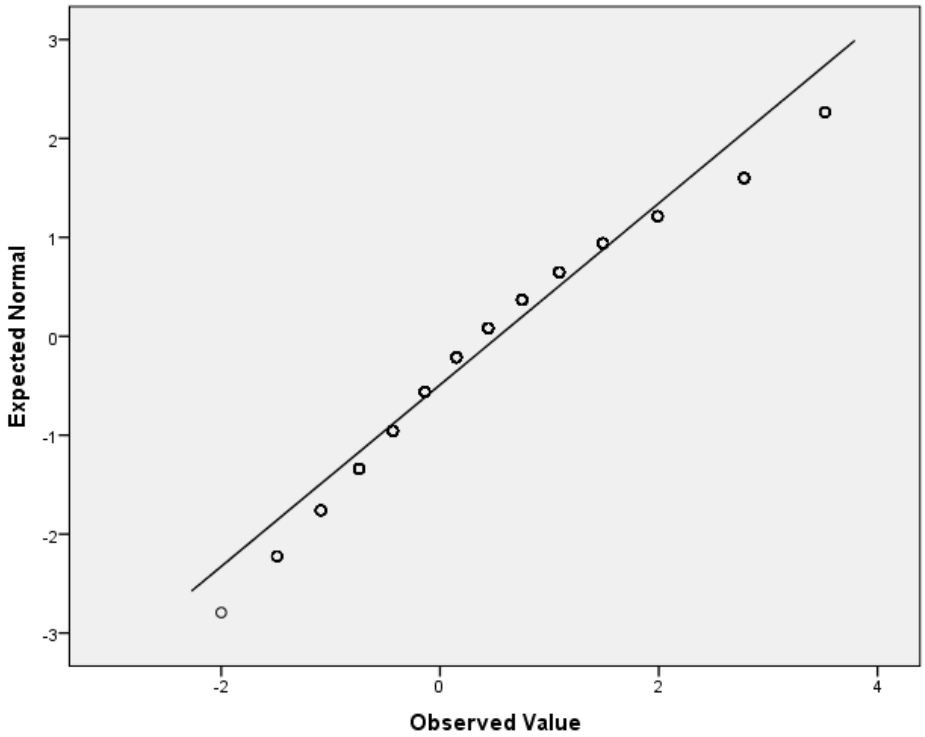

Figure 270: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Lecture Test 1 2012

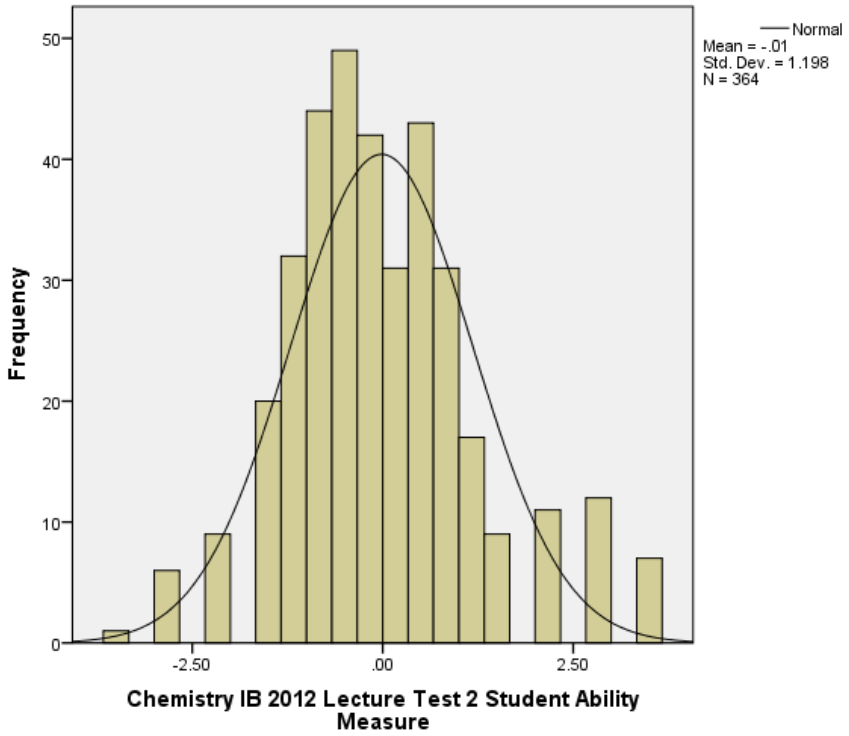

Figure 271: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Chemistry IB 2012 to Determine the Distribution that the Measures Follow

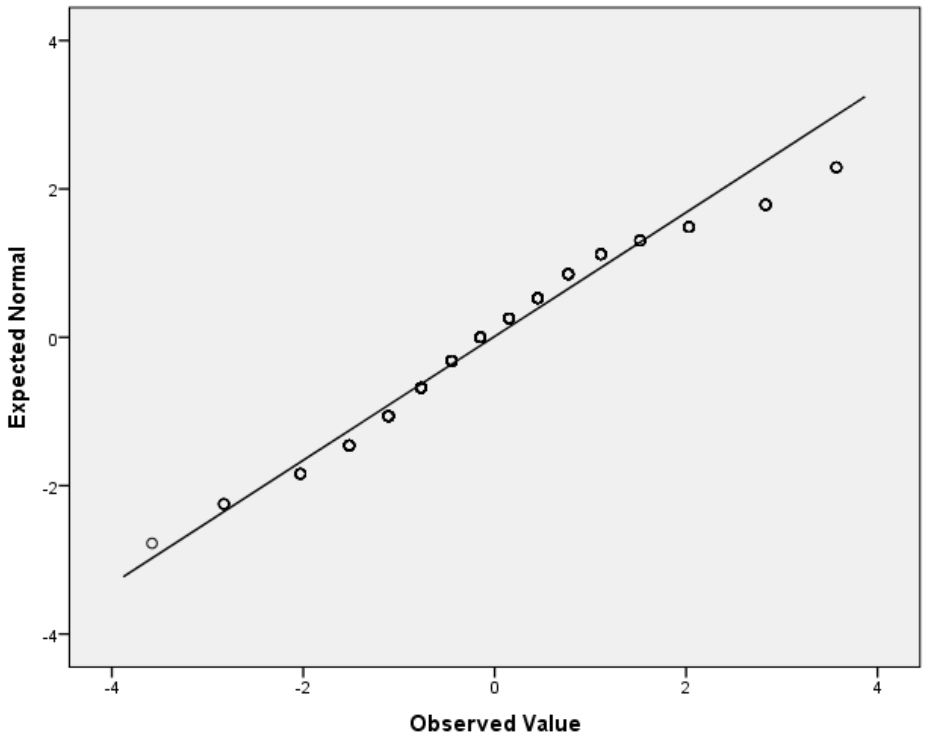

Figure 272: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Lecture Test 2 2012

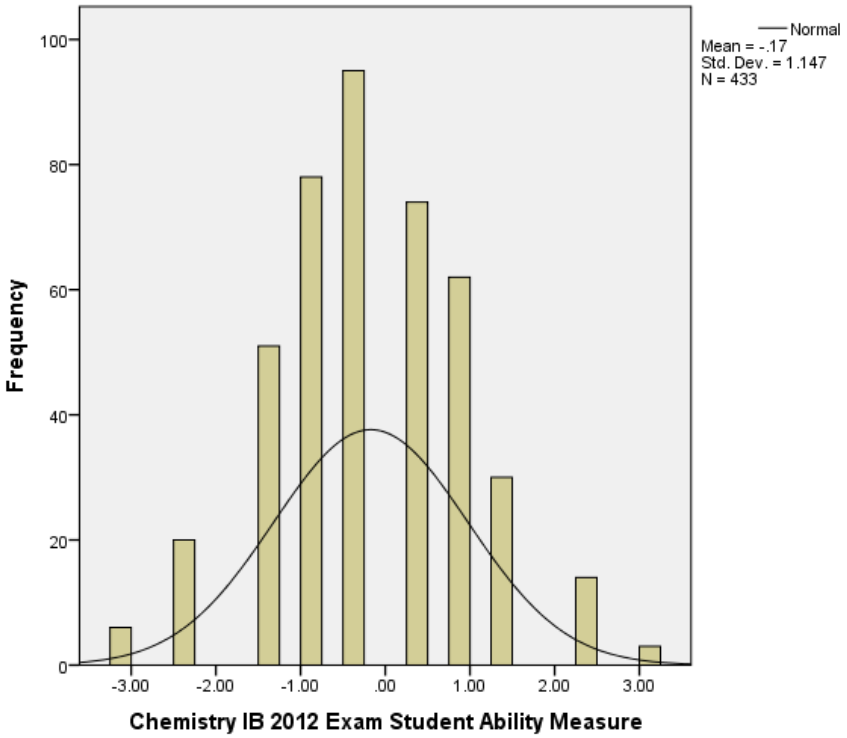

Figure 273: Histogram of the Rasch Student Ability Measures in Exam from Chemistry IB 2012 to Determine the Distribution that the Measures Follow

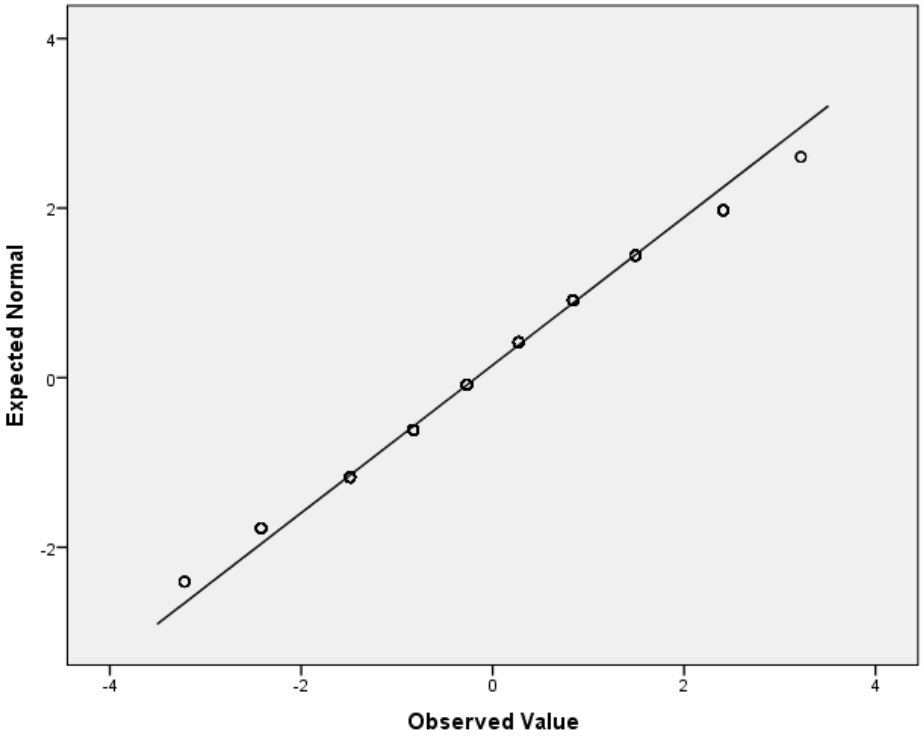

Figure 274: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Exam 2012

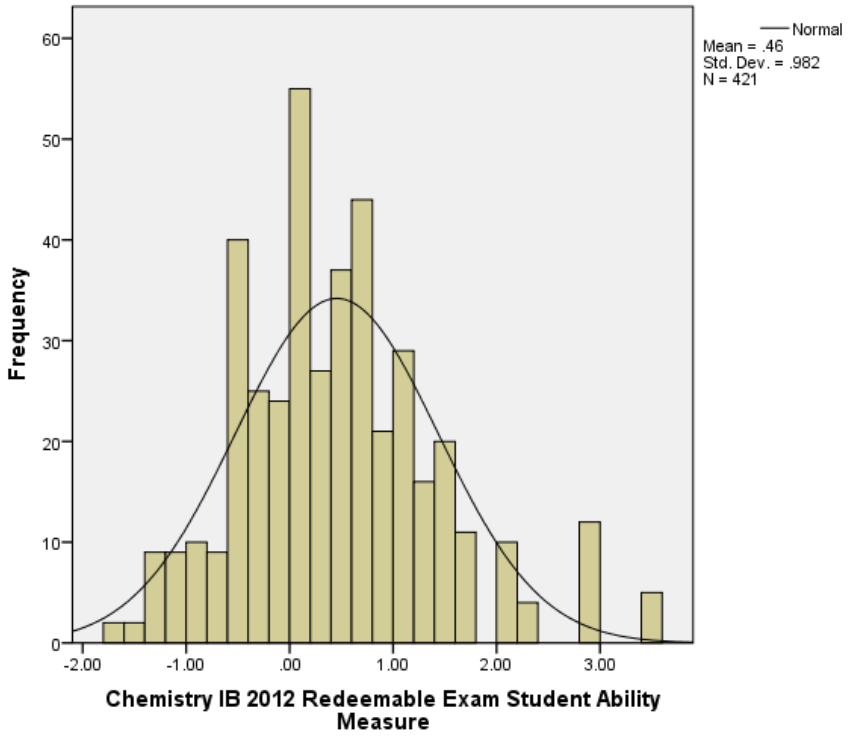

Figure 275: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Chemistry IB 2012 to Determine the Distribution that the Measures Follow

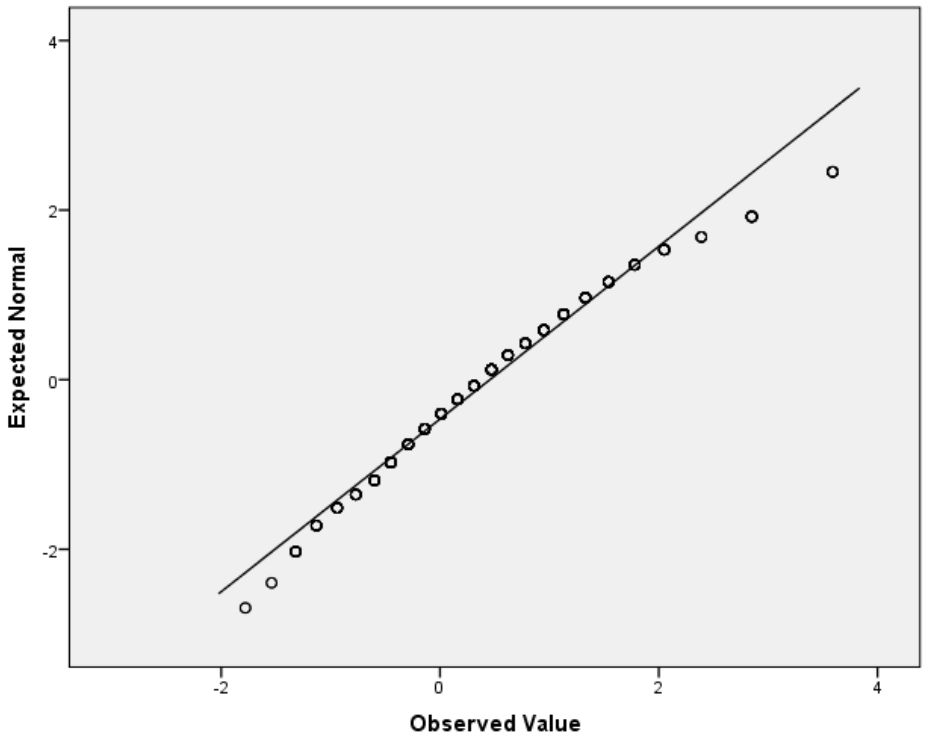

Figure 276: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Redeemable Exam 2012

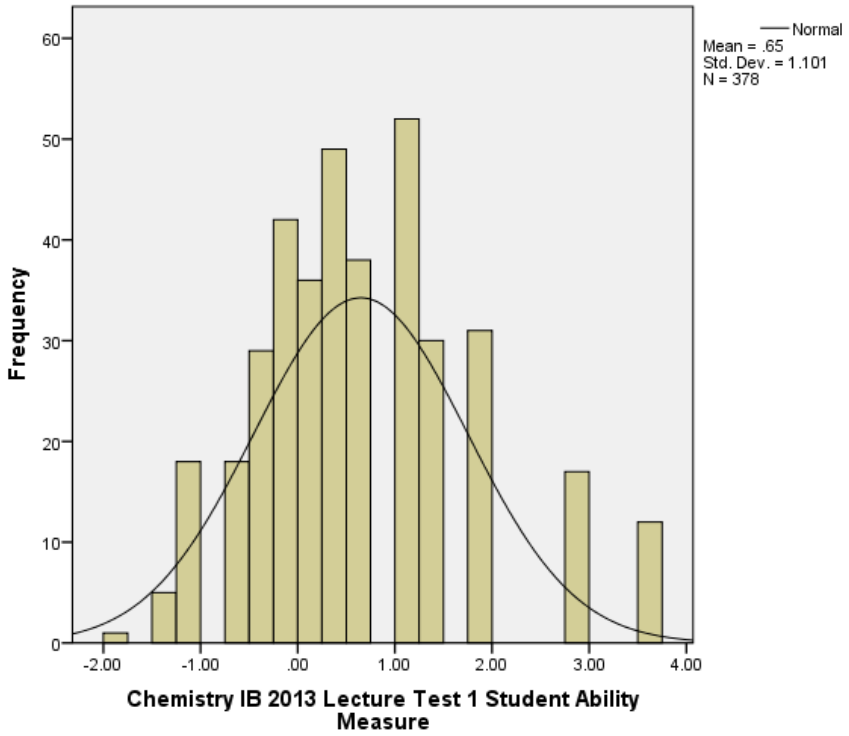

Figure 277: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IB 2013 to Determine the Distribution that the Measures Follow

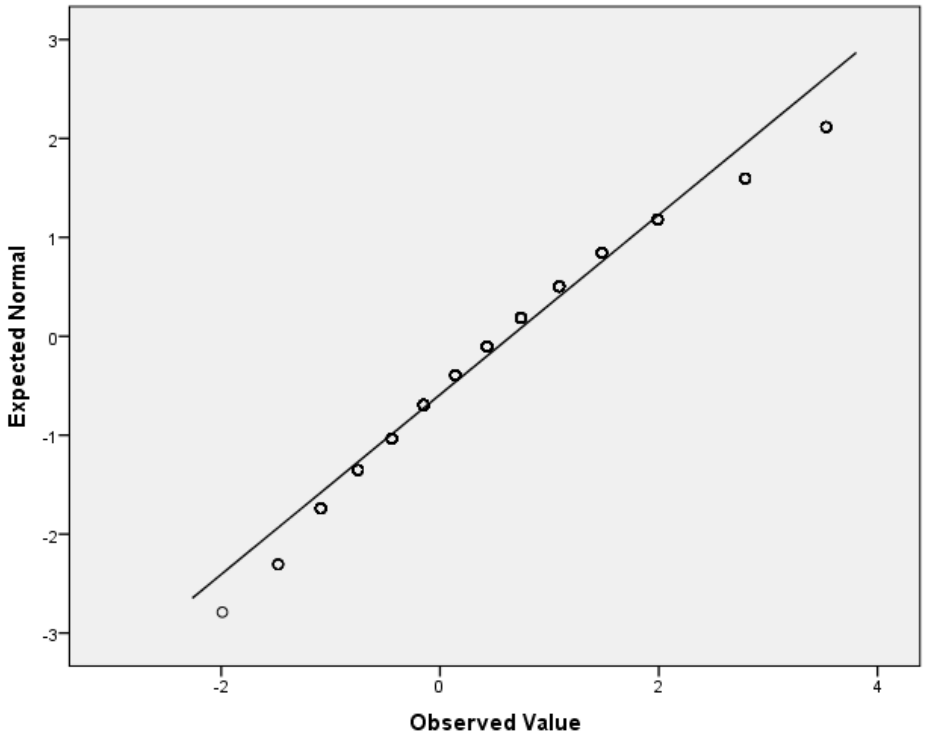

Figure 278: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Lecture Test 1 2013

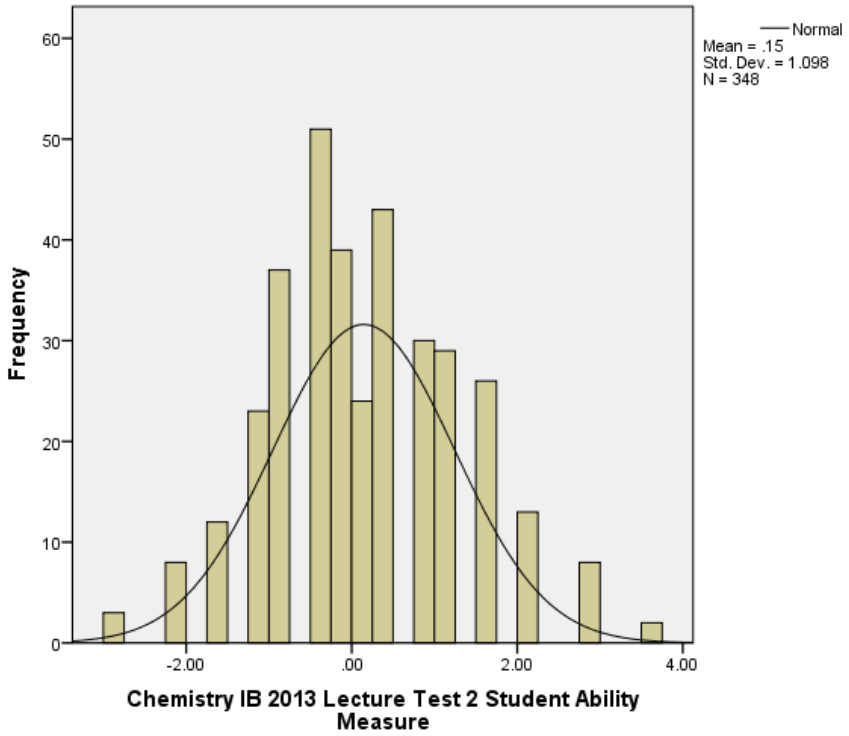

Figure 279: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Chemistry IB 2013 to Determine the Distribution that the Measures Follow

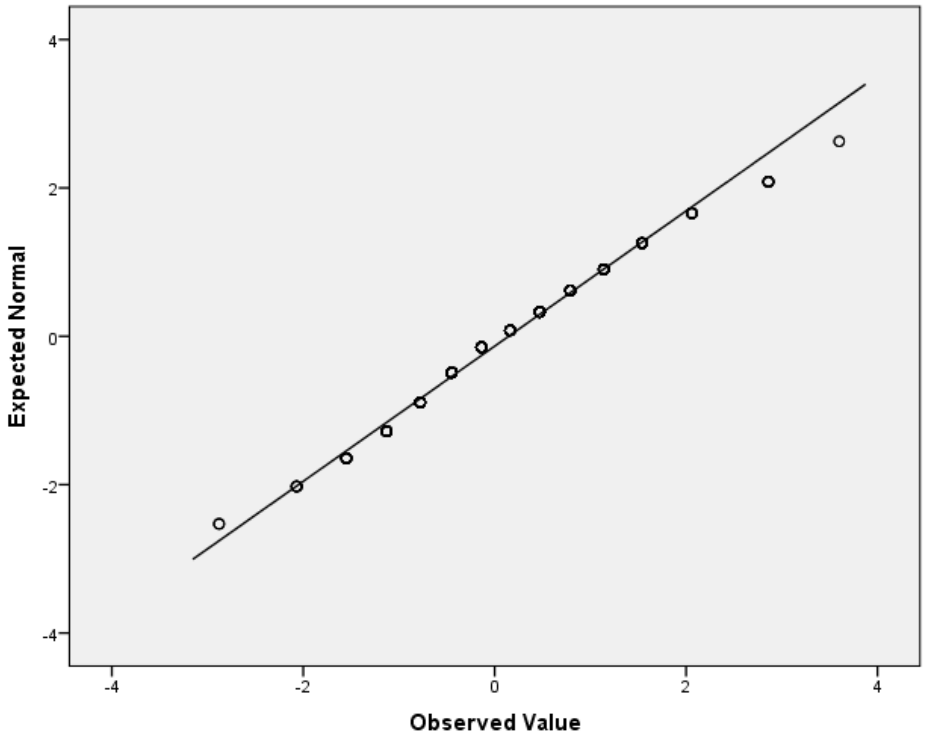

Figure 280: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Lecture Test 2 2013

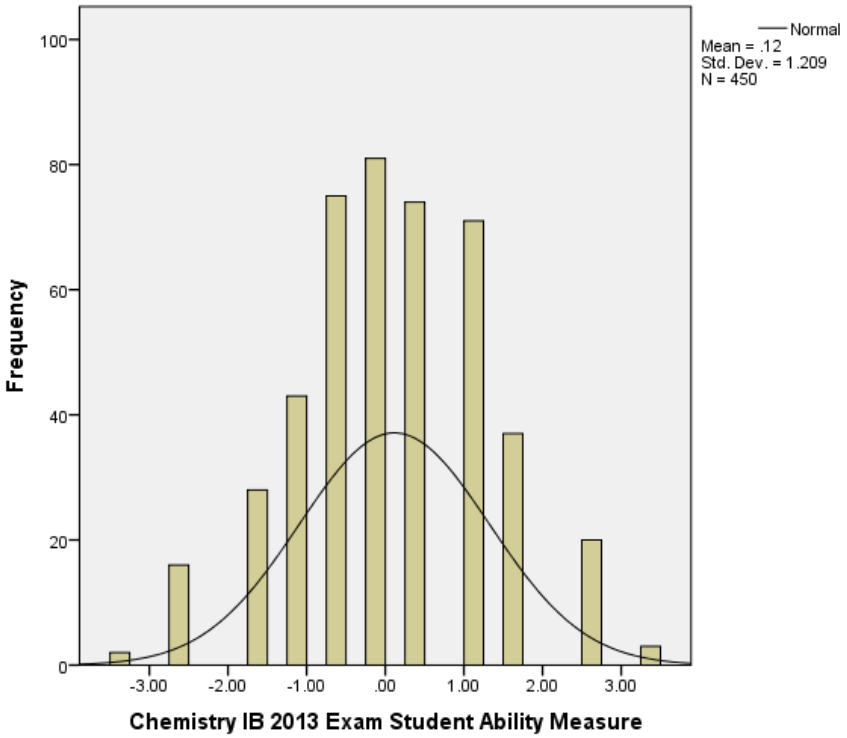

Figure 281: Histogram of the Rasch Student Ability Measures in Exam from Chemistry IB 2013 to Determine the Distribution that the Measures Follow

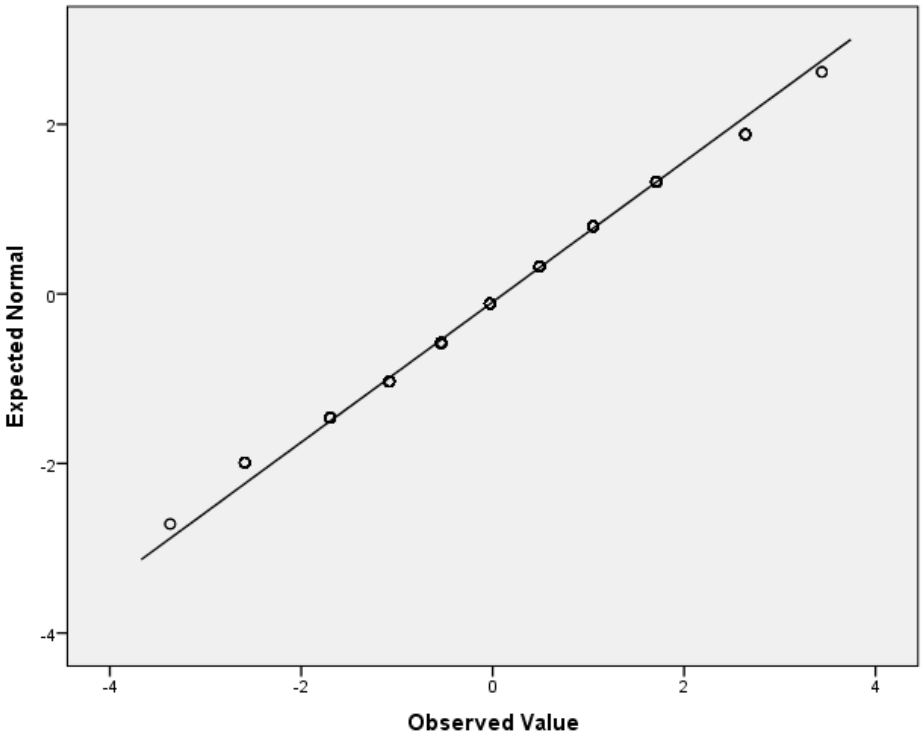

Figure 282: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Exam 2013

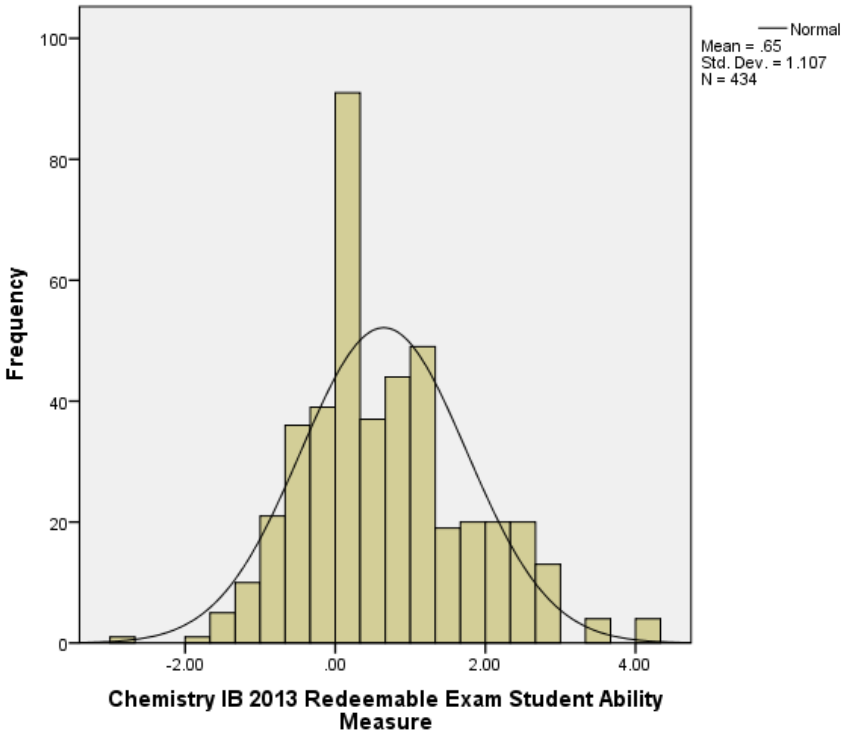

Figure 283: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Chemistry IB 2013 to Determine the Distribution that the Measures Follow

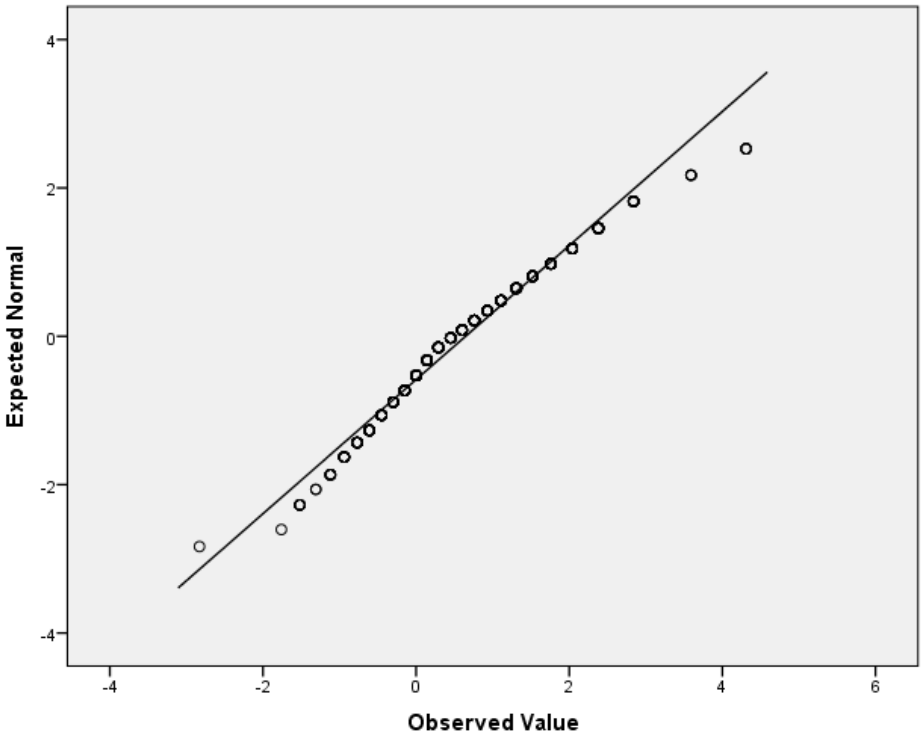

Figure 284: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Redeemable Exam 2013

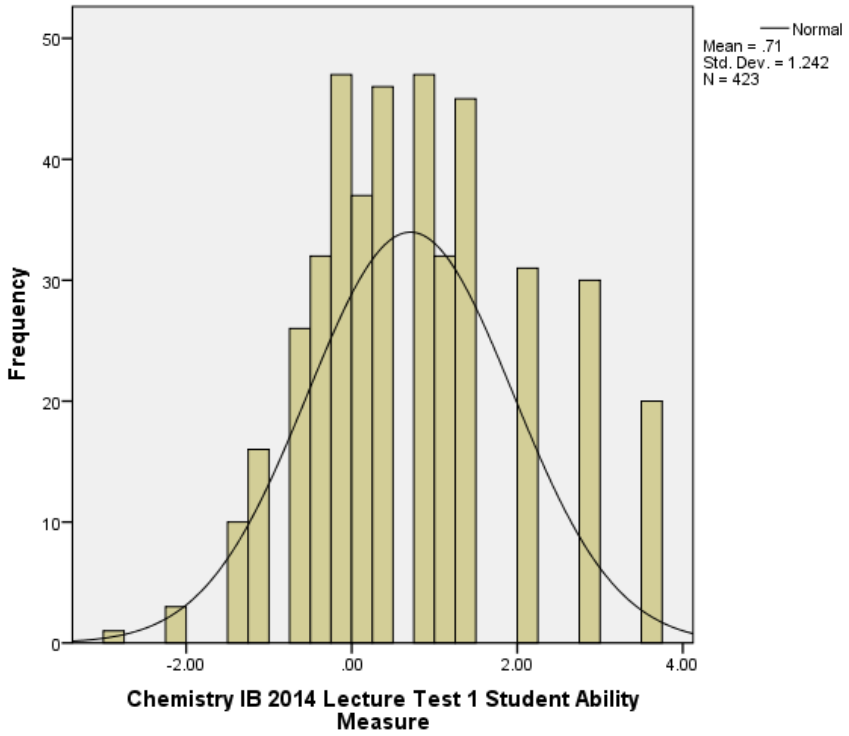

Figure 285: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IB 2014 to Determine the Distribution that the Measures Follow

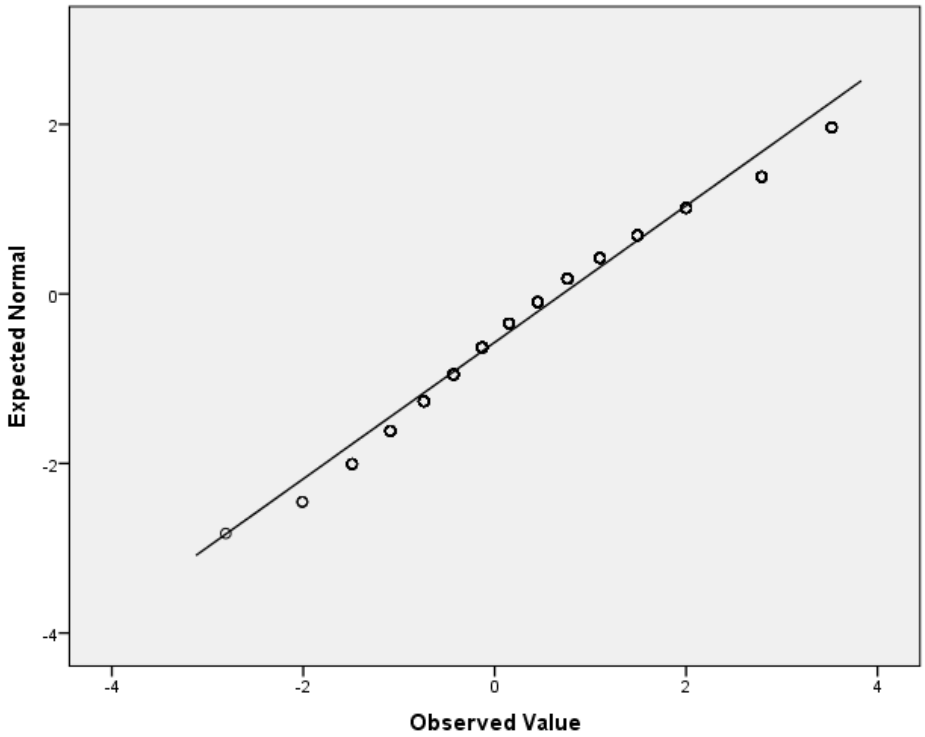

Figure 286: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Lecture Test 1 2014

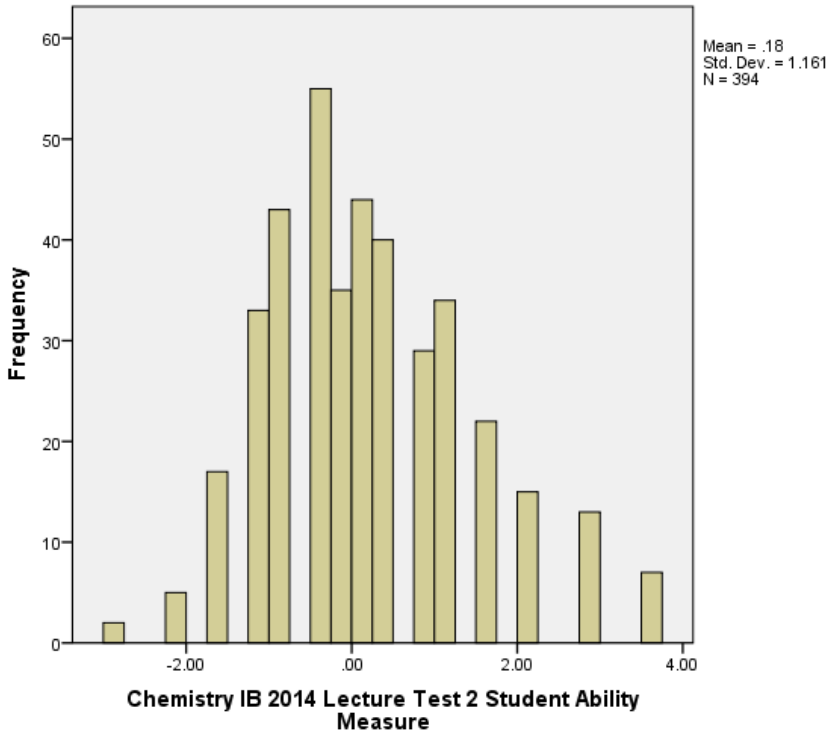

Figure 287: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Chemistry IB 2014 to Determine the Distribution that the Measures Follow

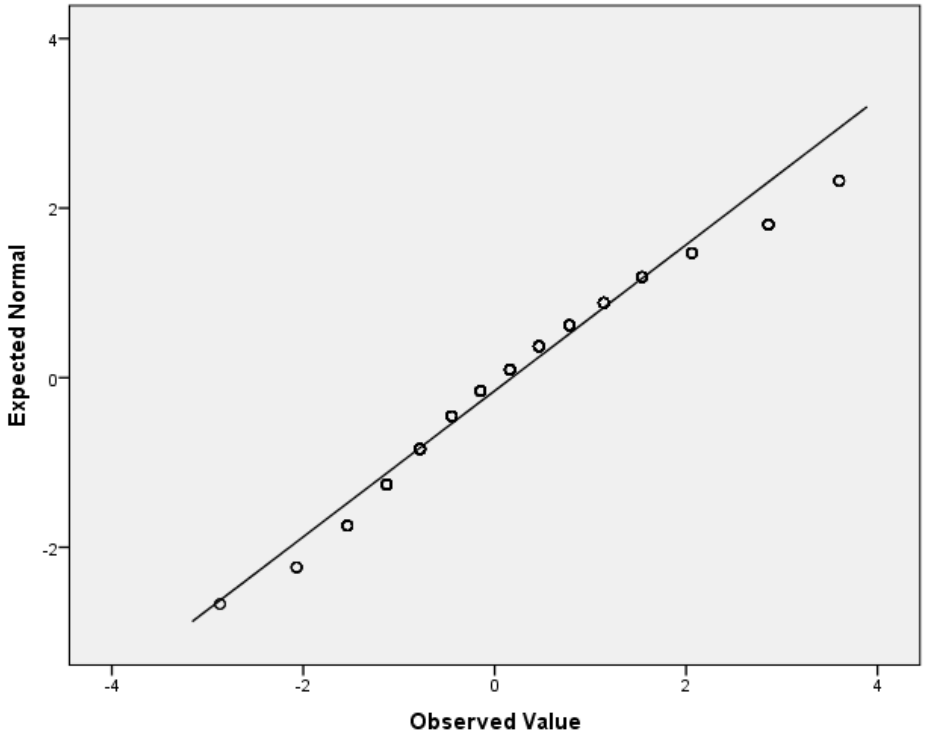

Figure 288: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Lecture Test 2 2014

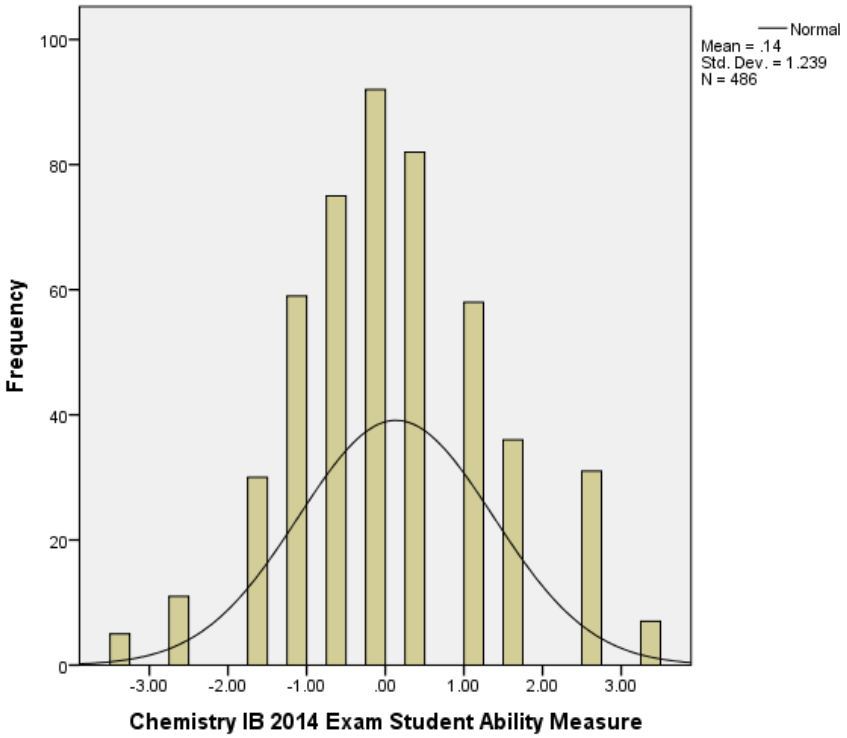

Figure 289: Histogram of the Rasch Student Ability Measures in Exam from Chemistry IB 2014 to Determine the Distribution that the Measures Follow

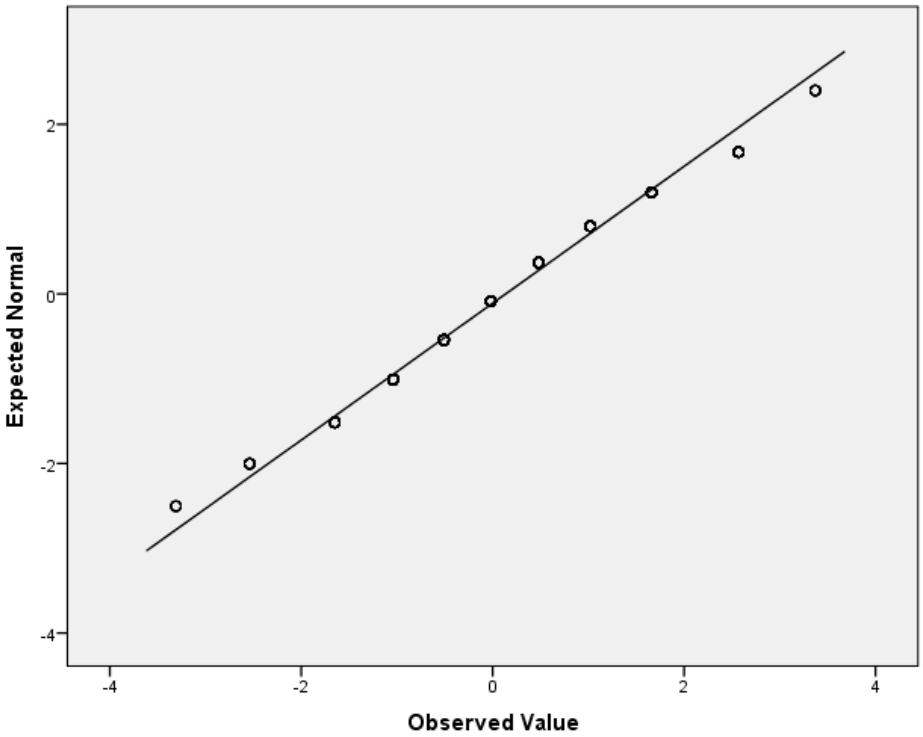

Figure 290: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Exam 2014

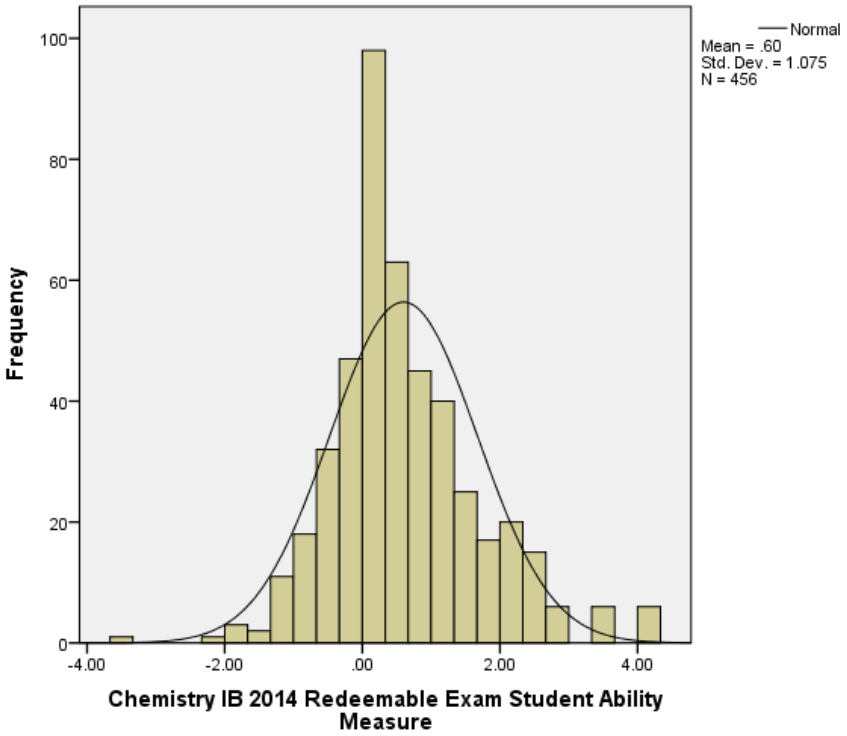

Figure 291: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Chemistry IB 2014 to Determine the Distribution that the Measures Follow

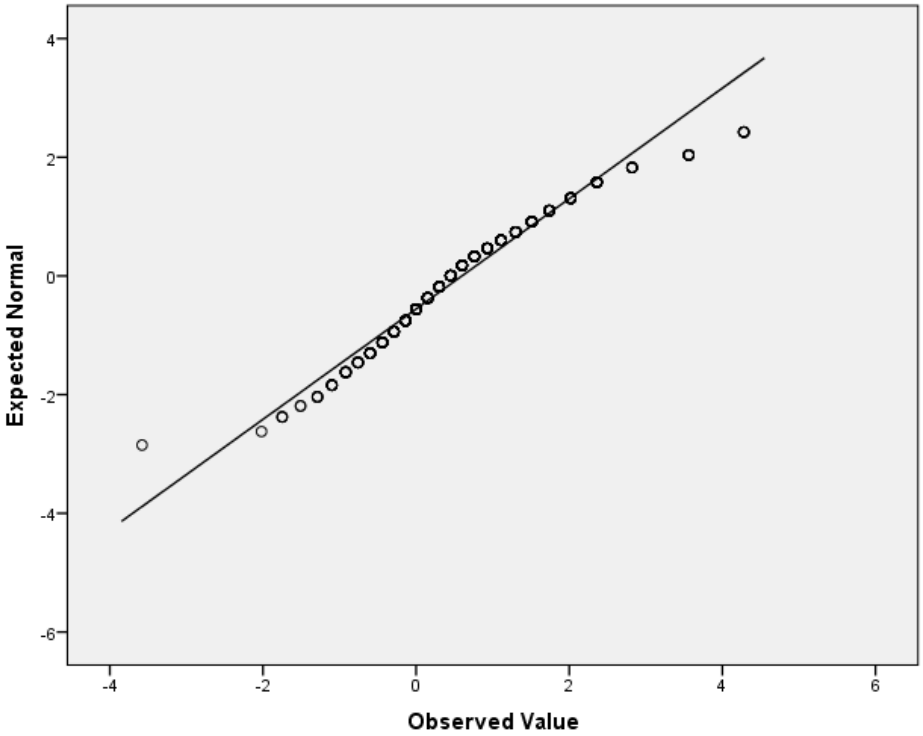

Figure 292: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Redeemable Exam 2014

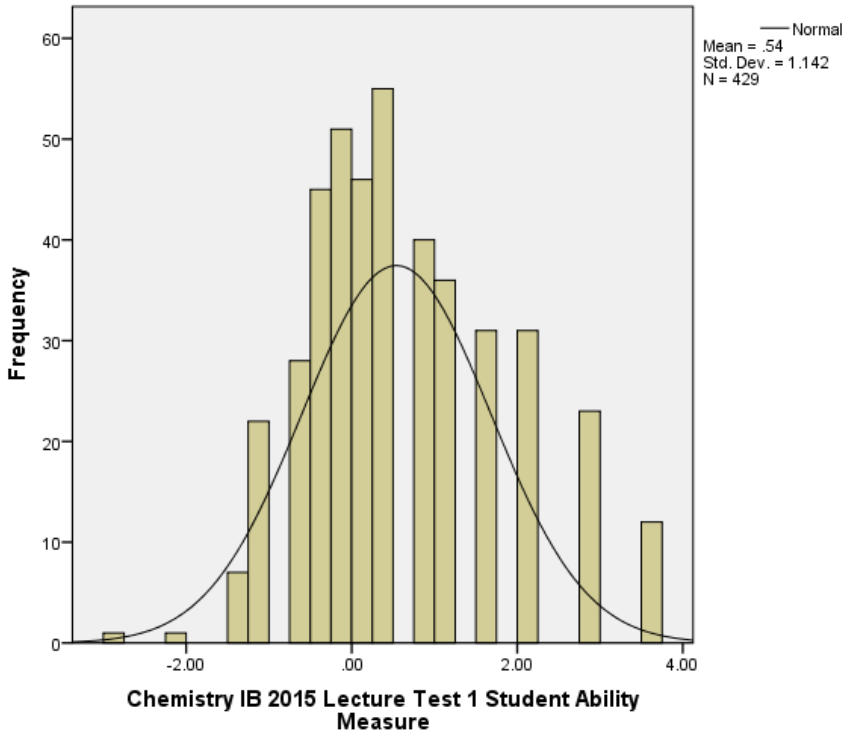

Figure 293: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Chemistry IB 2015 to Determine the Distribution that the Measures Follow

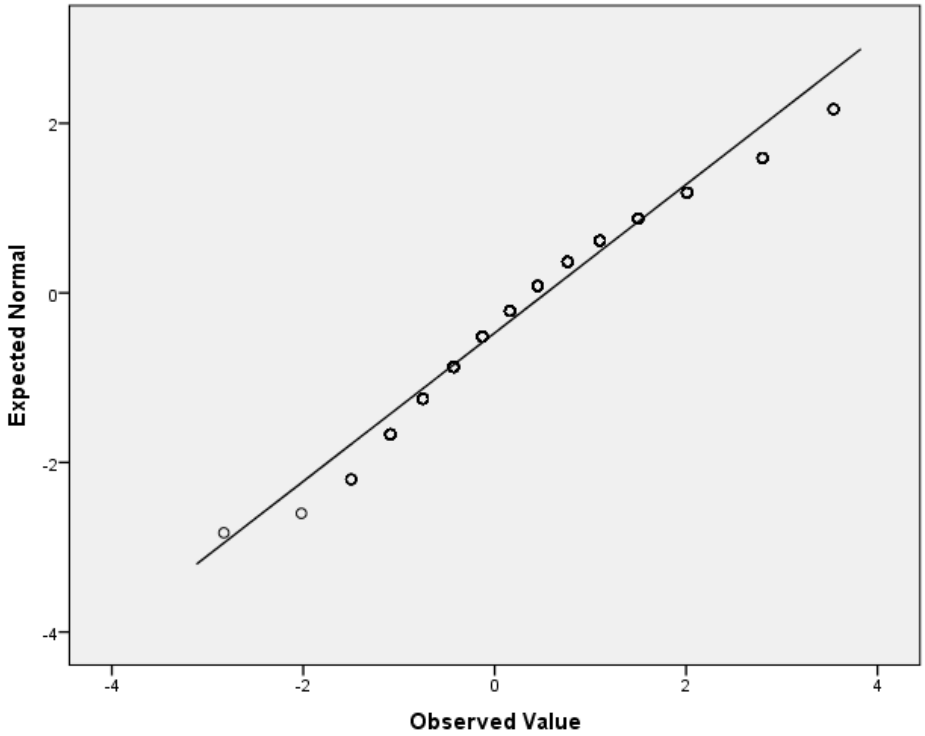

Figure 294: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Lecture Test 1 2015

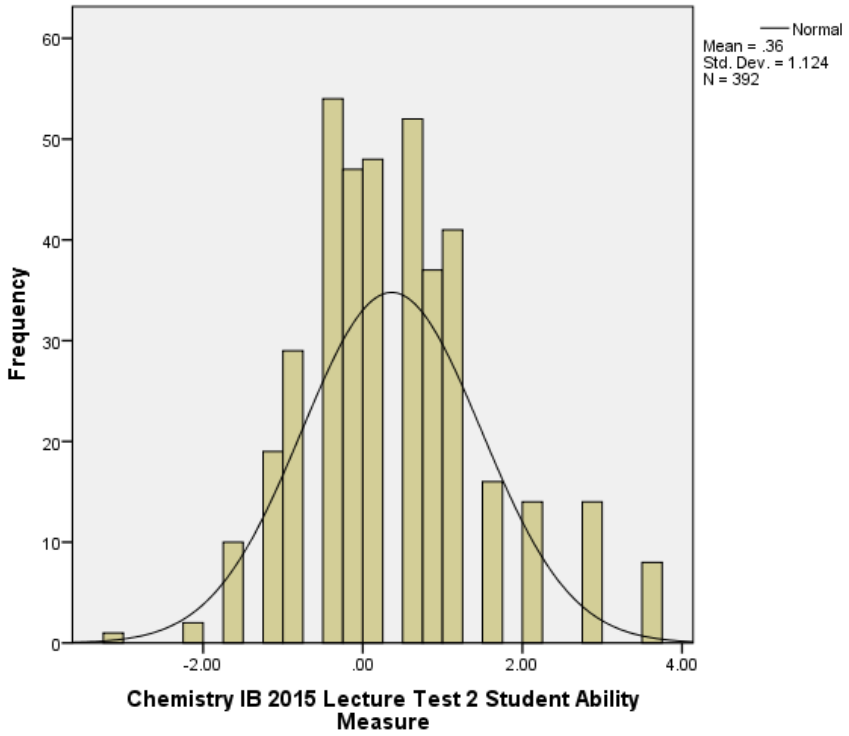

Figure 295: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Chemistry IB 2015 to Determine the Distribution that the Measures Follow

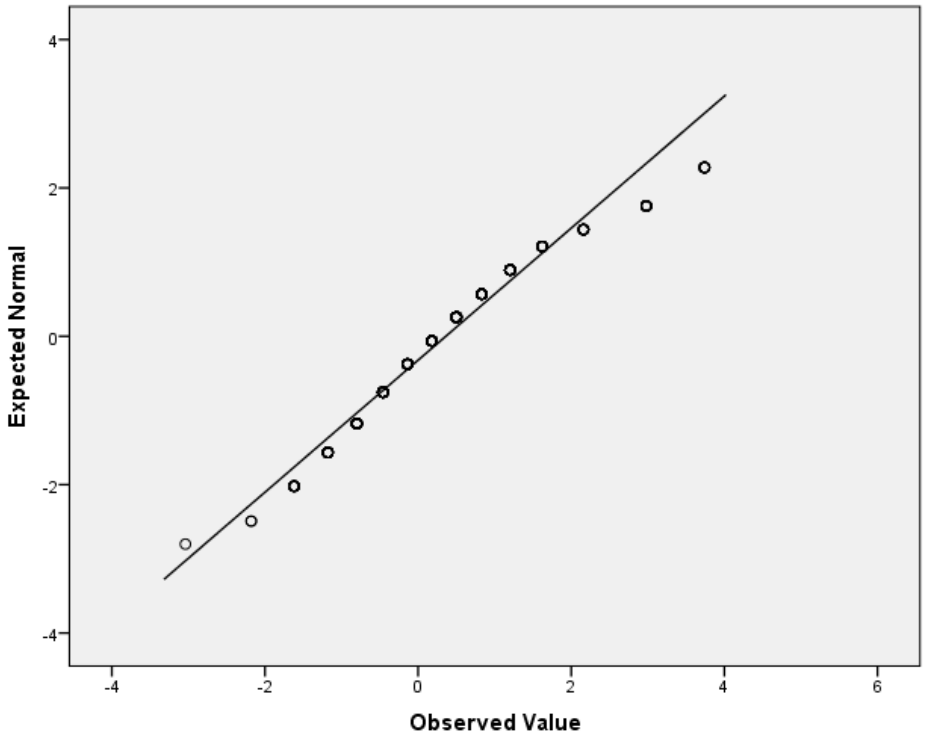

Figure 296: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Lecture Test 2 2015

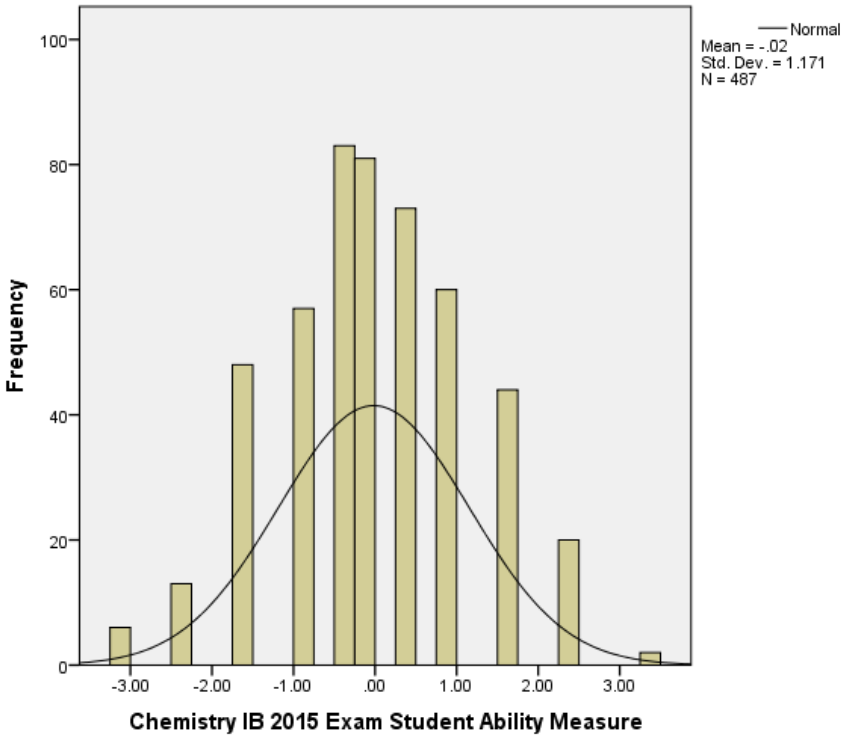

Figure 297: Histogram of the Rasch Student Ability Measures in Exam from Chemistry IB 2015 to Determine the Distribution that the Measures Follow

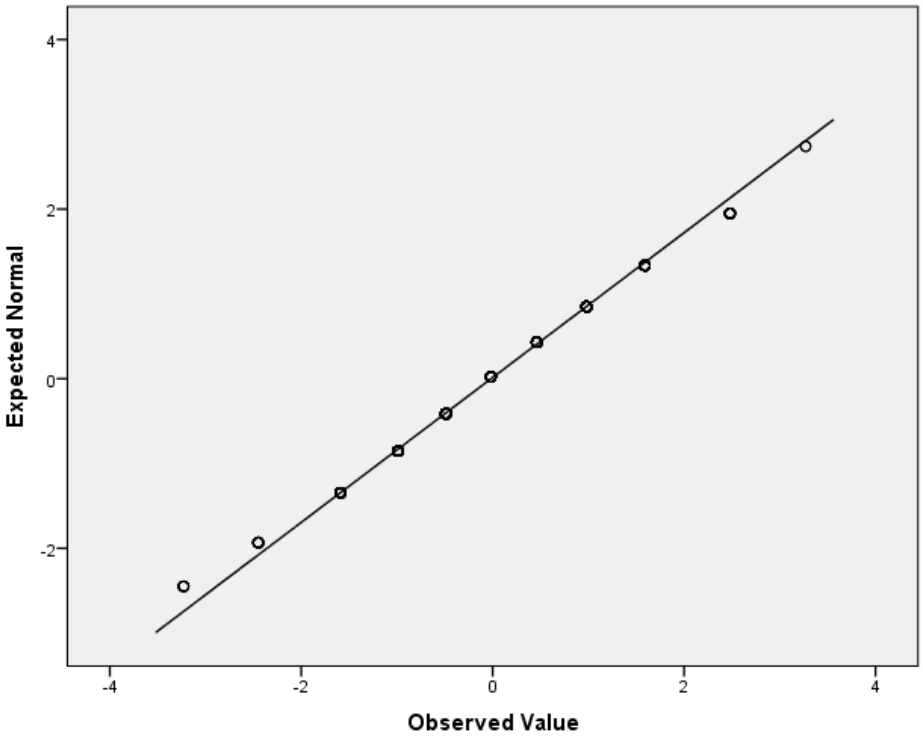

Figure 298: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Exam 2015

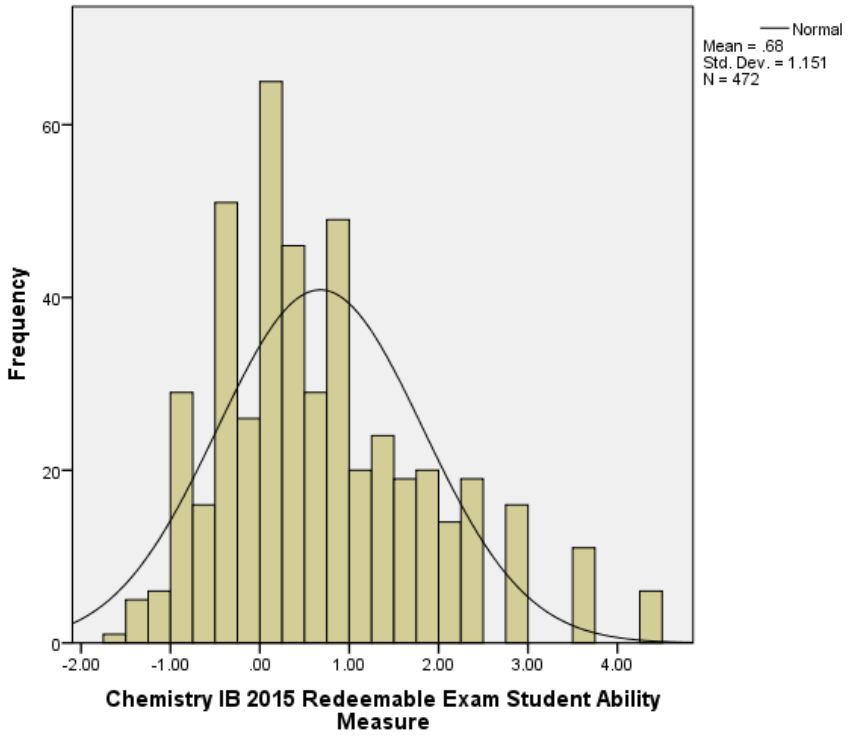

Figure 299: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Chemistry IB 2015 to Determine the Distribution that the Measures Follow

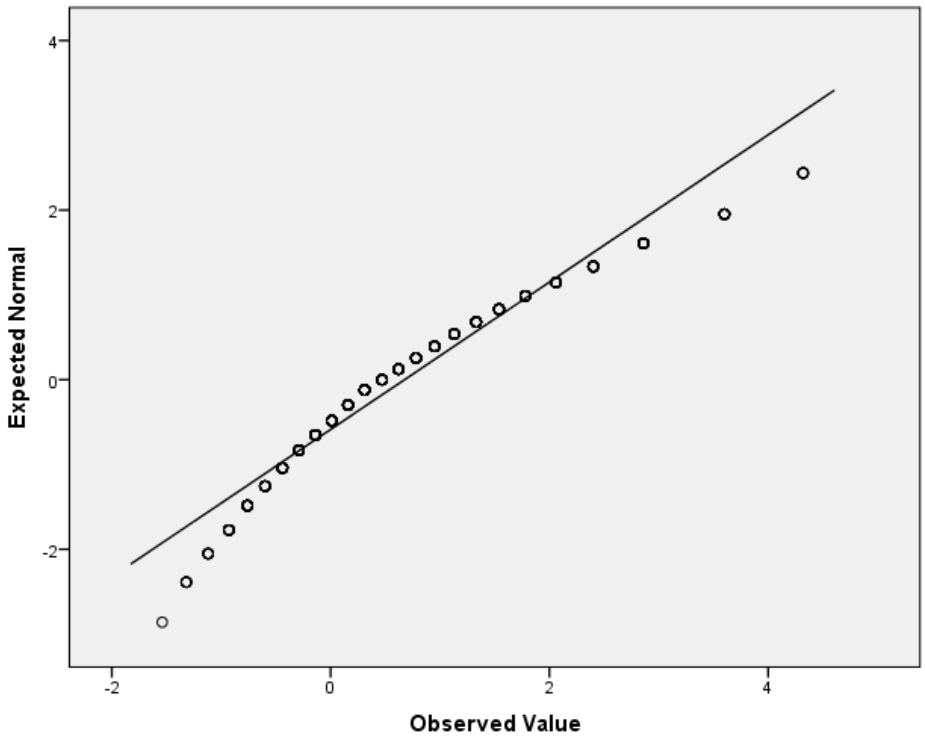

Figure 300: Rasch Student Ability Measure Q-Q Plot from Chemistry IB Redeemable Exam 2015

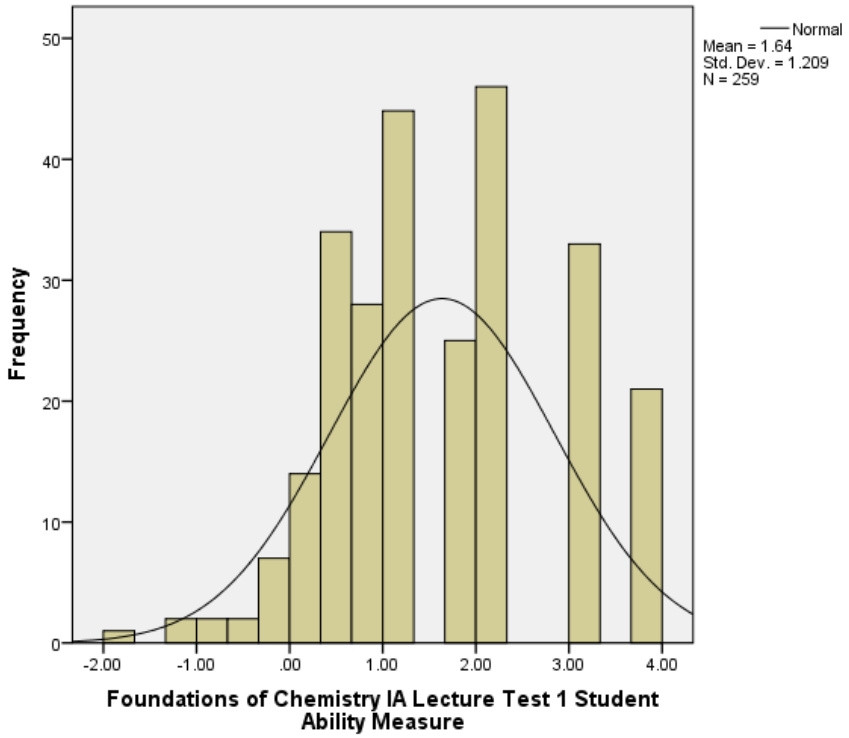

Figure 301: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Foundations of Chemistry IA 2012 to Determine the Distribution that the Measures Follow

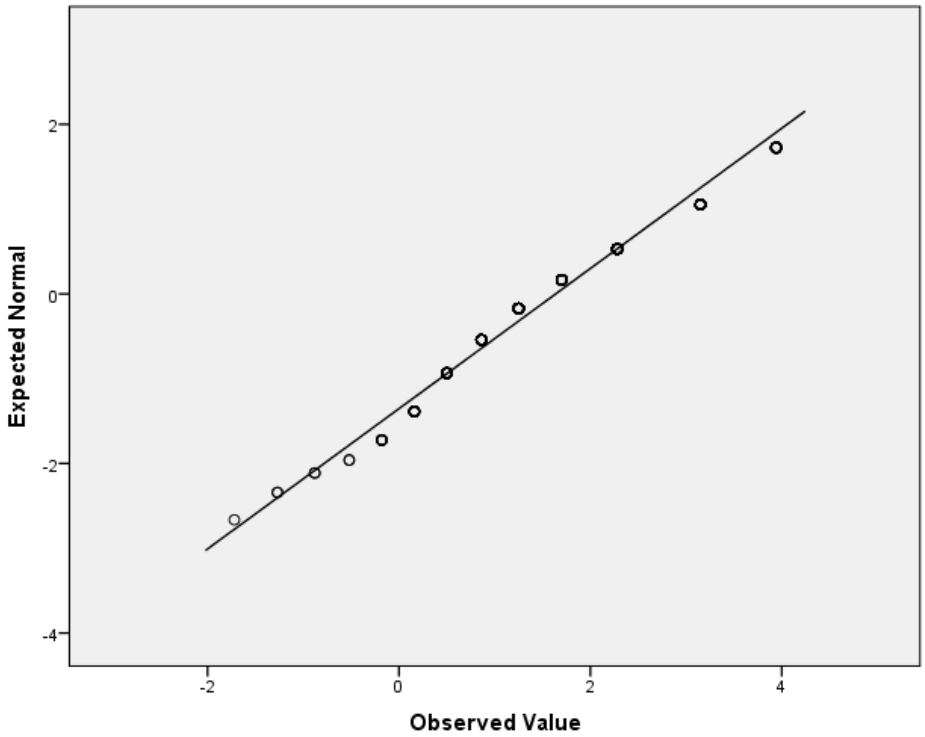

Figure 302: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Lecture Test 1 2012

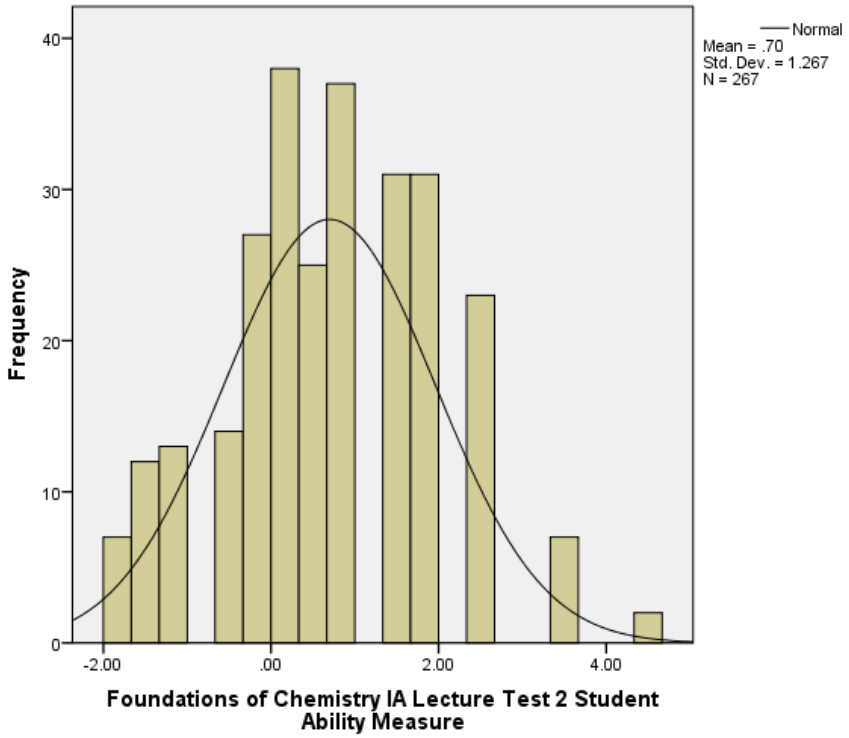

Figure 303: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Foundations of Chemistry IA 2012 to Determine the Distribution that the Measures Follow

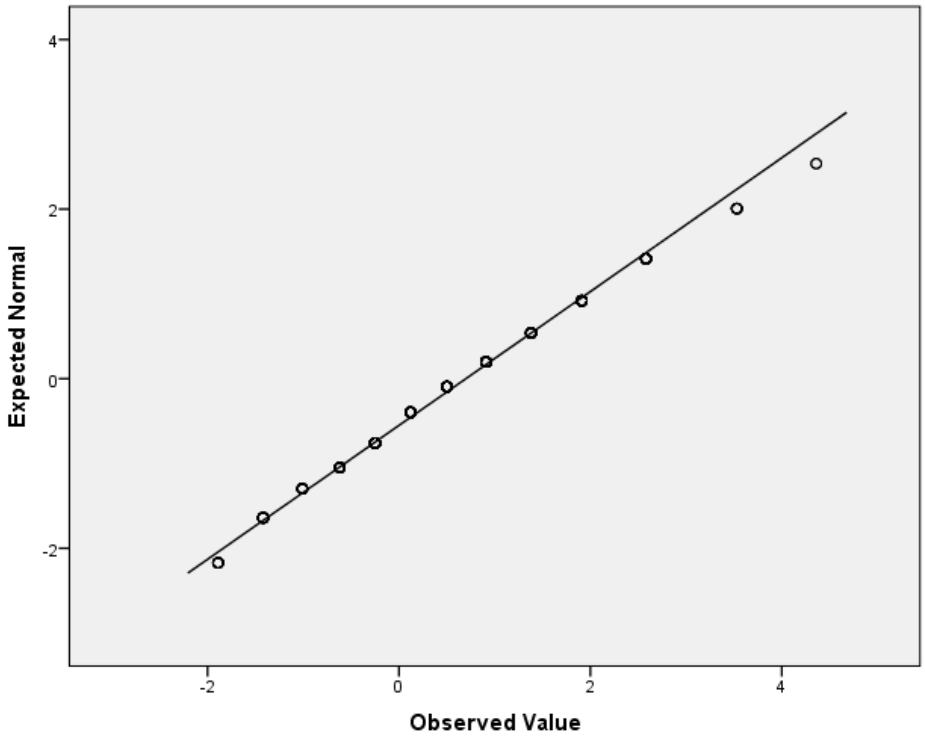

Figure 304: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Lecture Test 2 2012

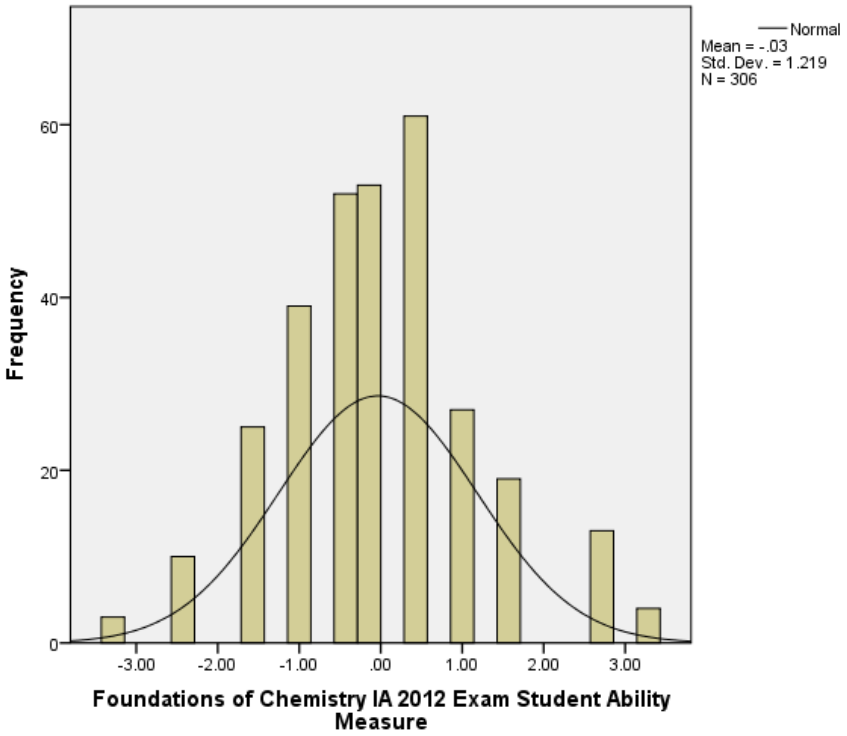

Figure 305: Histogram of the Rasch Student Ability Measures in Exam from Foundations of Chemistry IA 2012 to Determine the Distribution that the Measures Follow

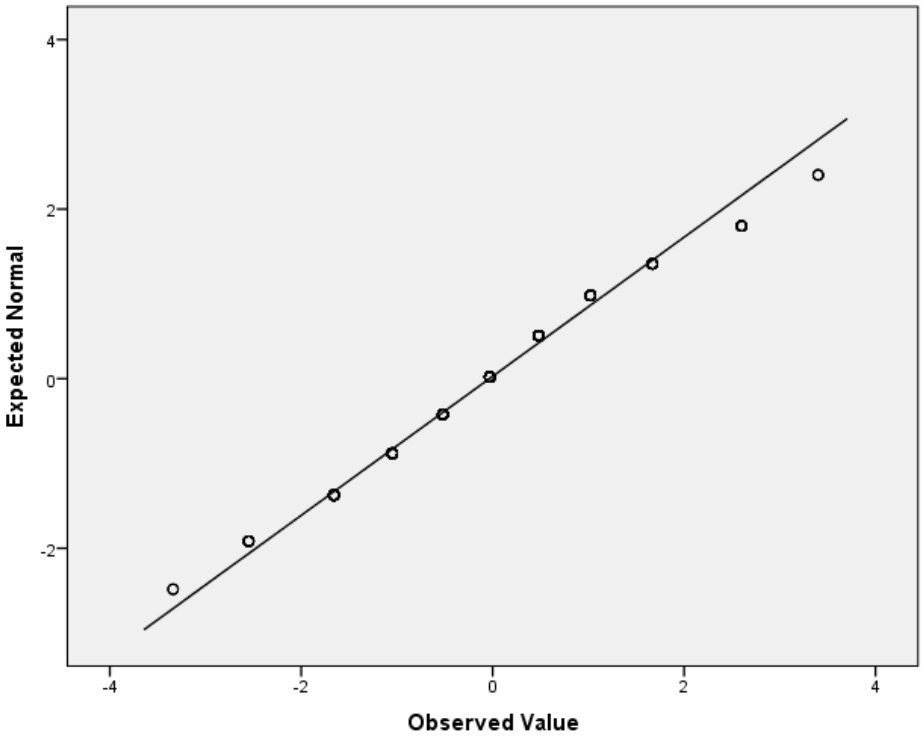

Figure 306: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Exam 2012

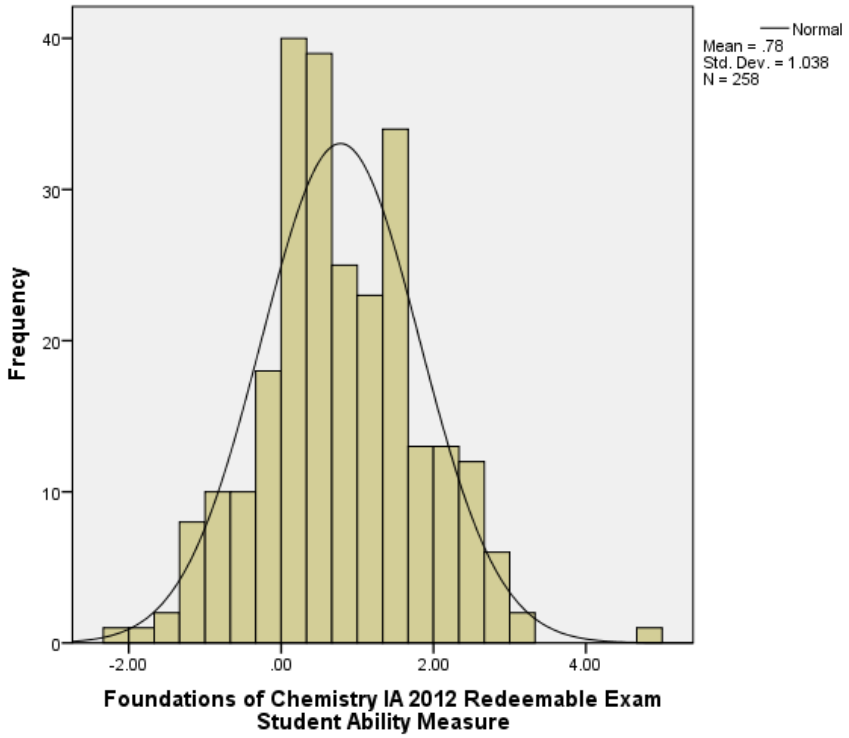

Figure 307: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Foundations of Chemistry IA 2012 to Determine the Distribution that the Measures Follow

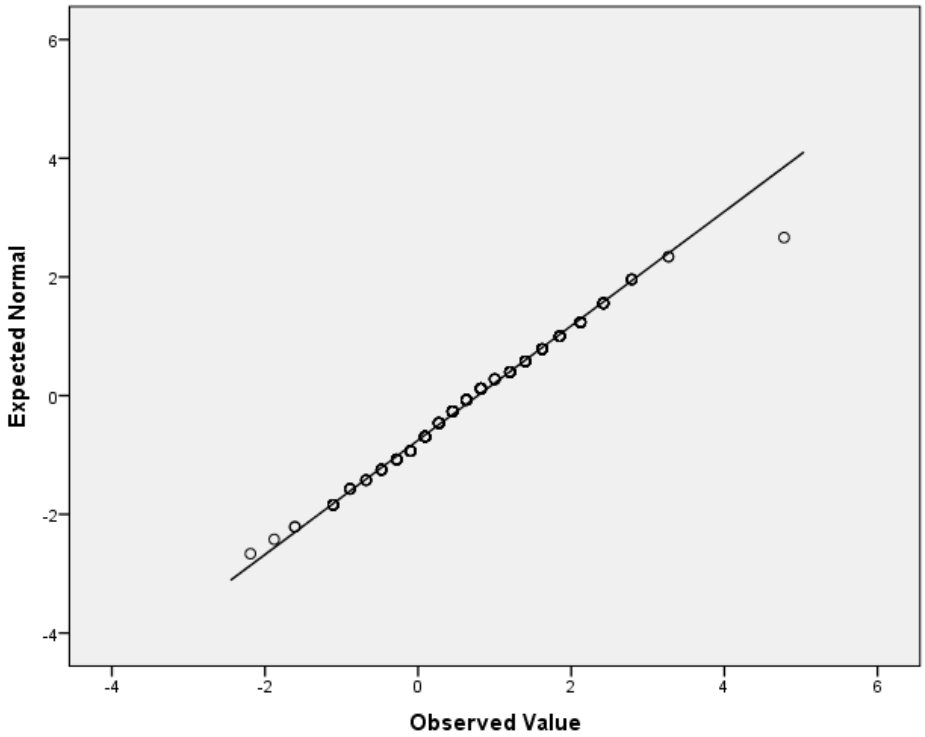

Figure 308: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Redeemable Exam 2012

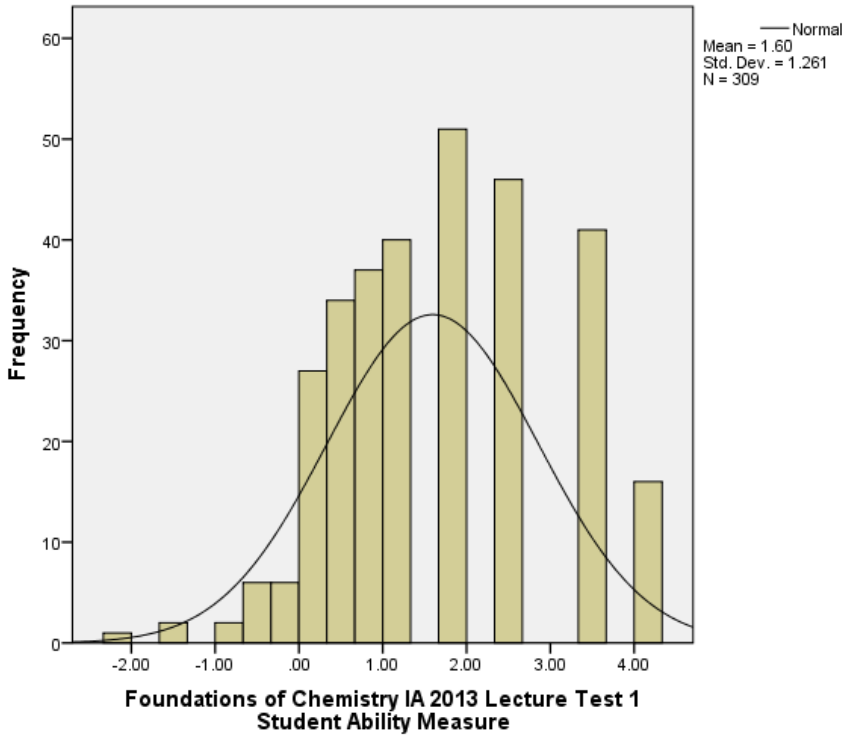

Figure 309: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Foundations of Chemistry IA 2013 to Determine the Distribution that the Measures Follow

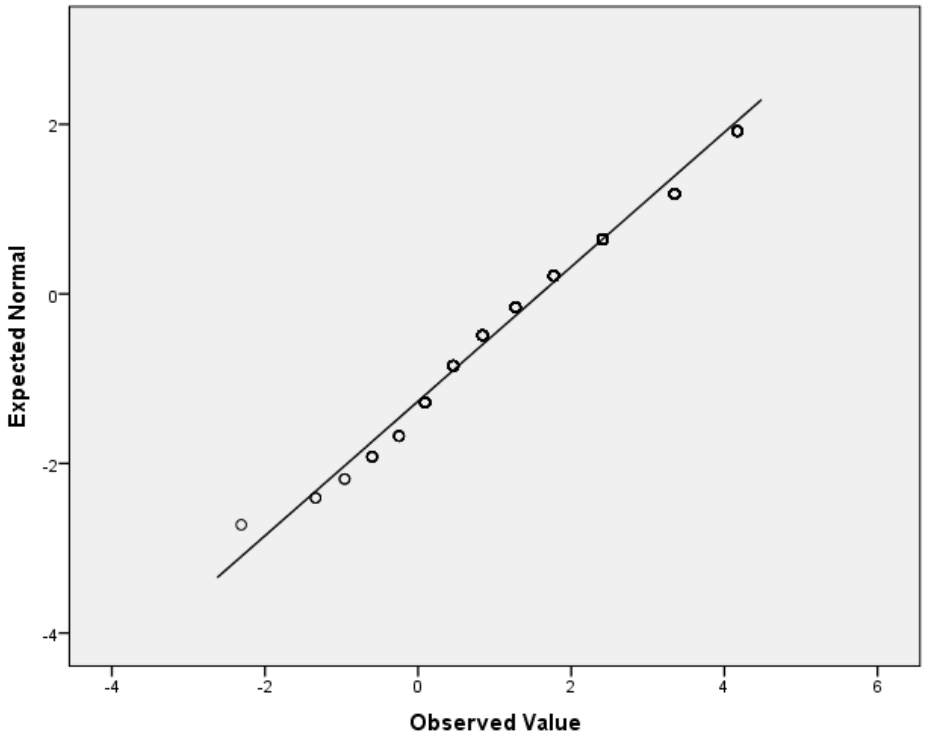

Figure 310: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Lecture Test 1 2013

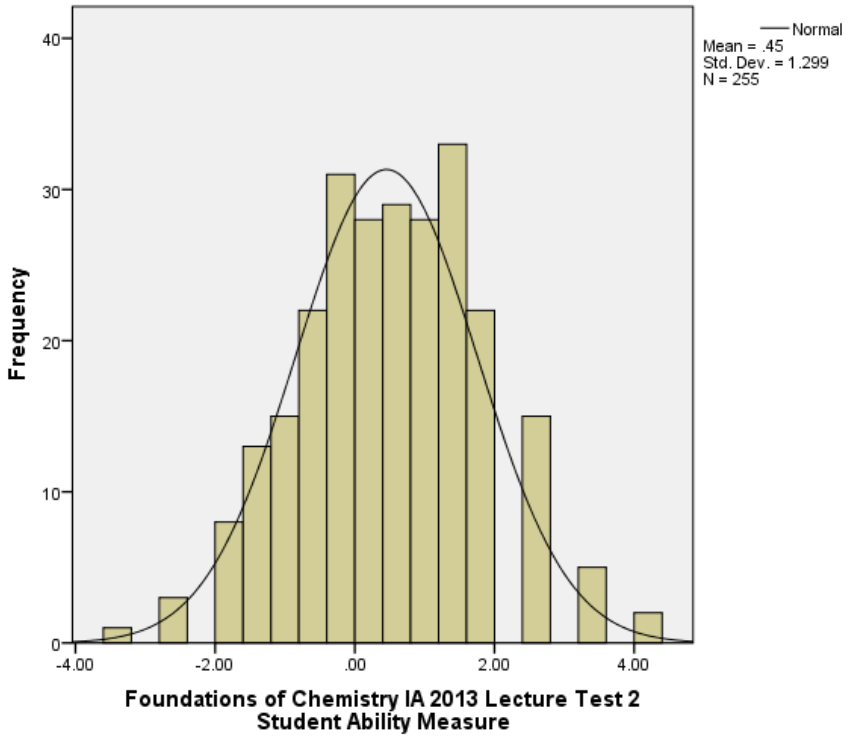

Figure 311: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Foundations of Chemistry IA 2013 to Determine the Distribution that the Measures Follow

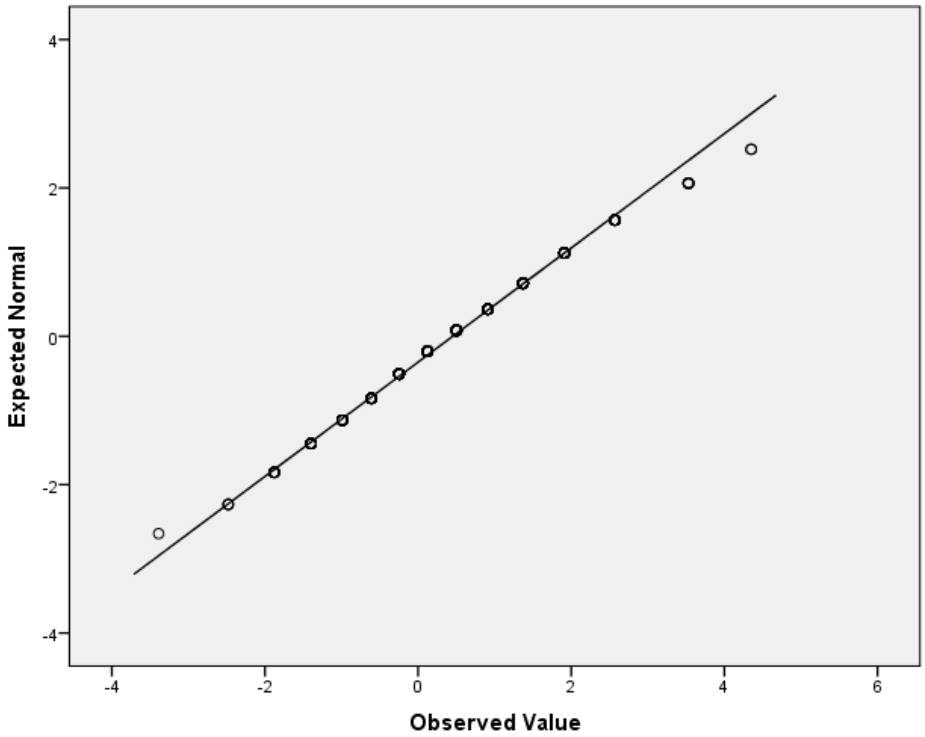

Figure 312: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Lecture Test 2 2013

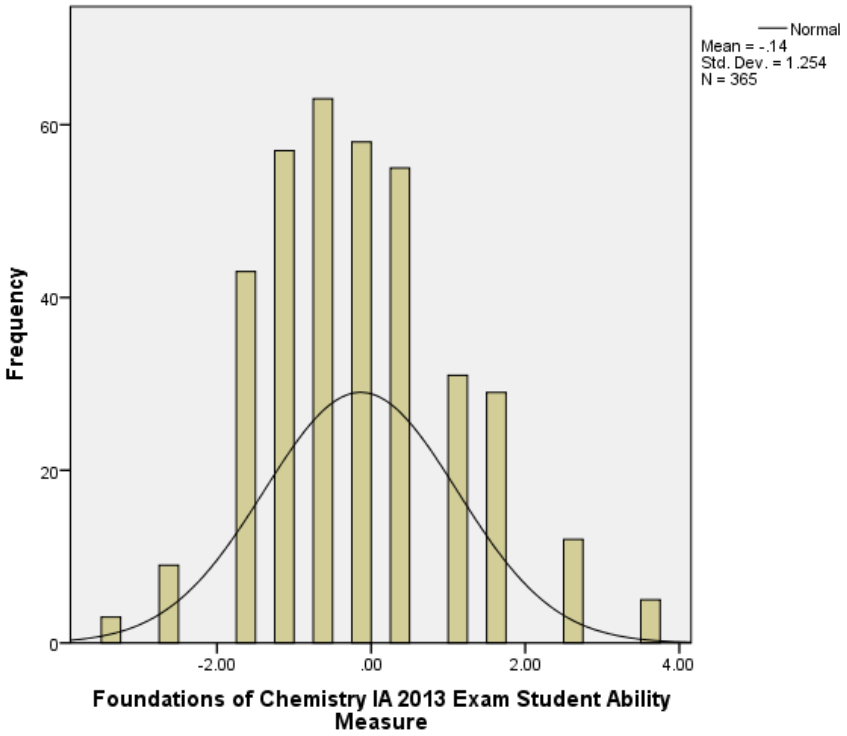

Figure 313: Histogram of the Rasch Student Ability Measures in Exam from Foundations of Chemistry IA 2013 to Determine the Distribution that the Measures Follow

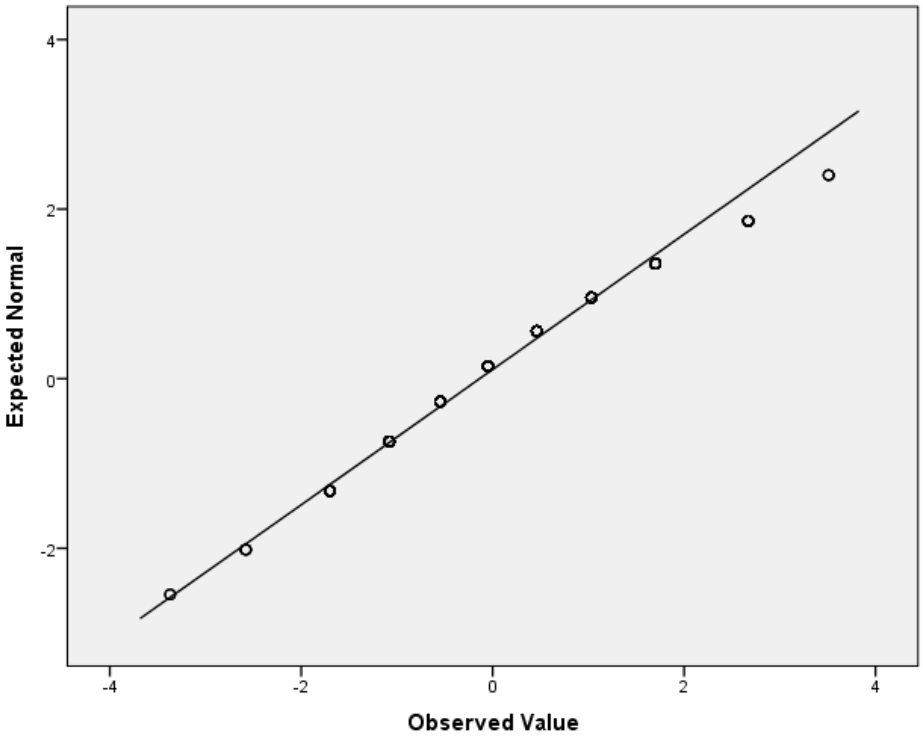

Figure 314: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Exam 2013

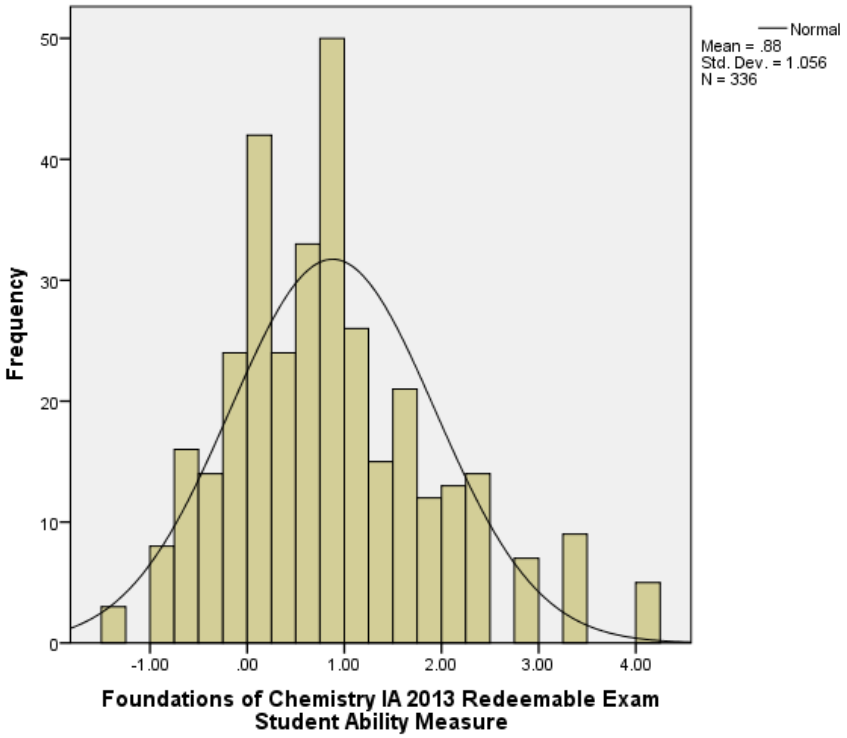

Figure 315: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Foundations of Chemistry IA 2013 to Determine the Distribution that the Measures Follow

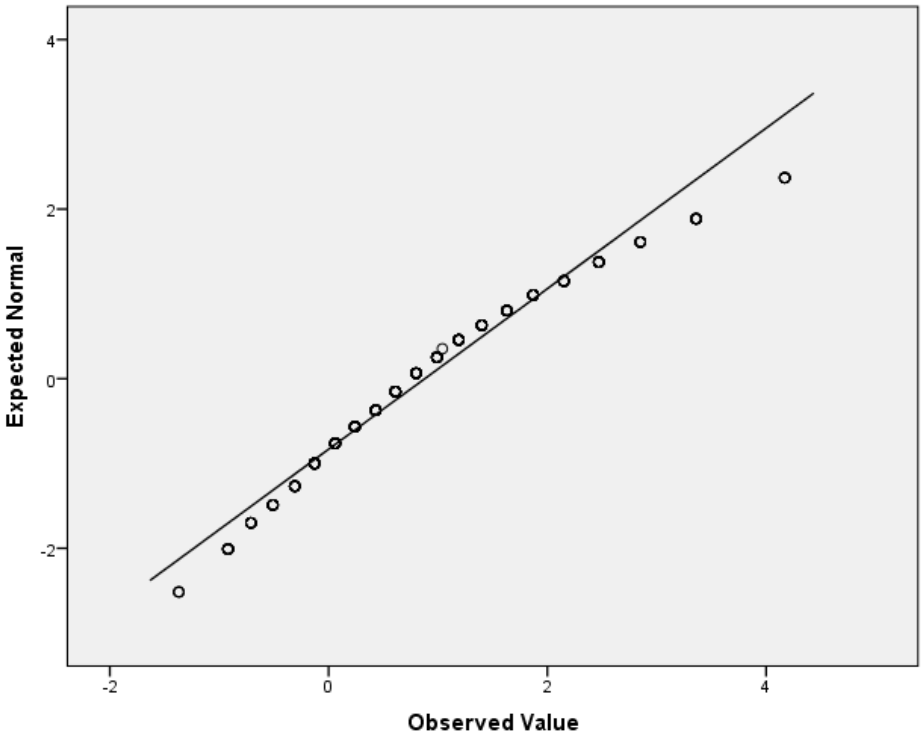

Figure 316: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Redeemable Exam 2013

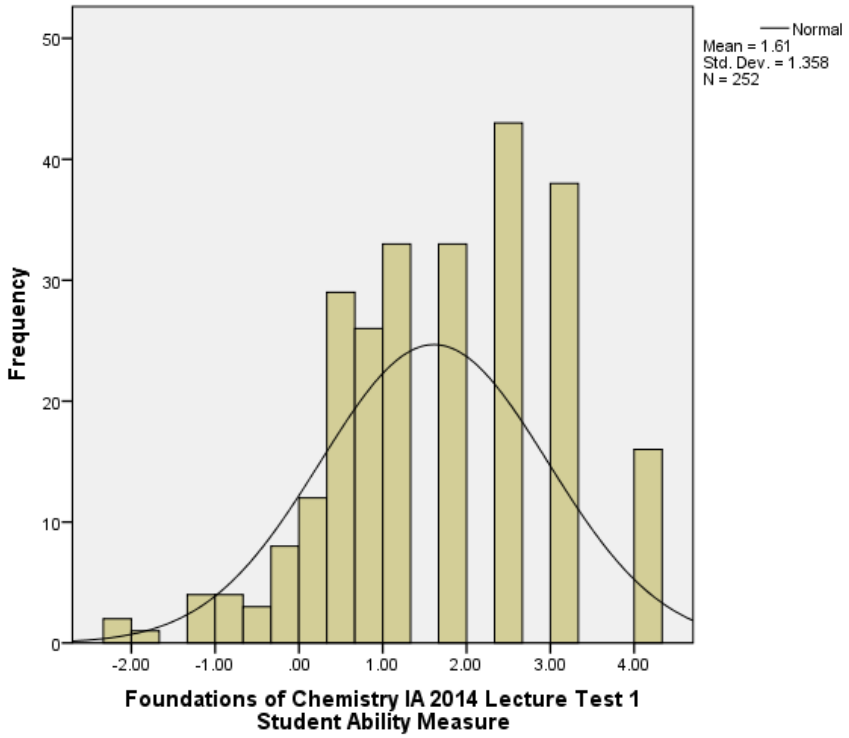

Figure 317: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Foundations of Chemistry IA 2014 to Determine the Distribution that the Measures Follow

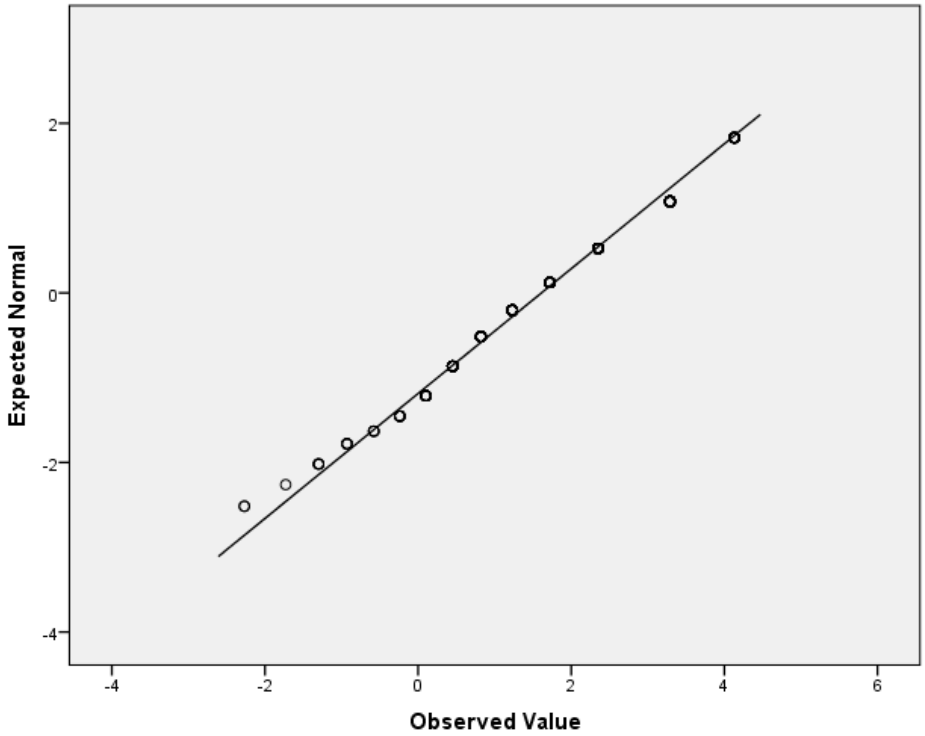

Figure 318: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Lecture Test 1 2014

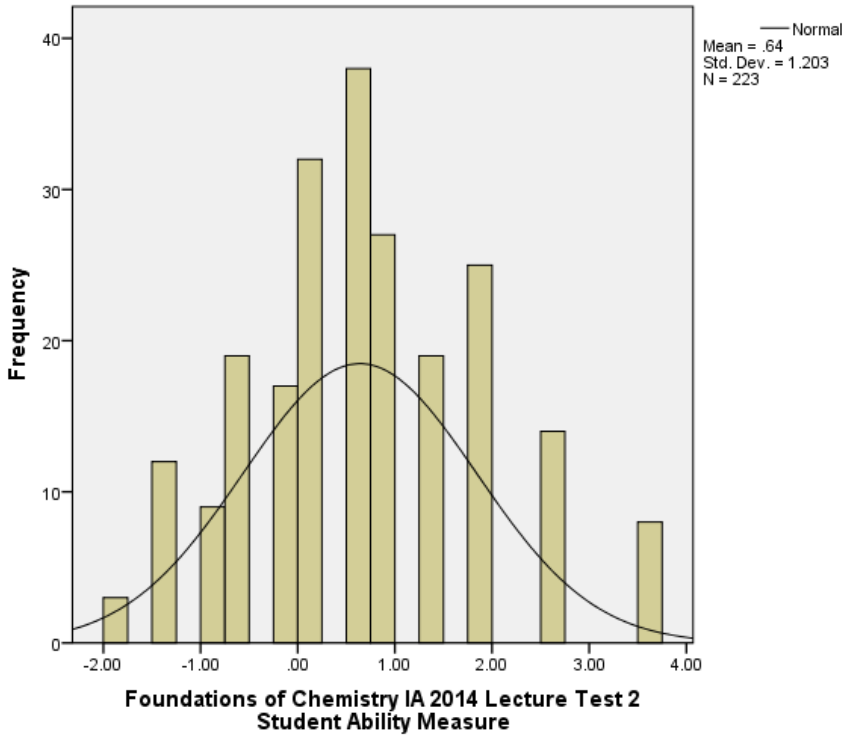

Figure 319: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Foundations of Chemistry IA 2014 to Determine the Distribution that the Measures Follow

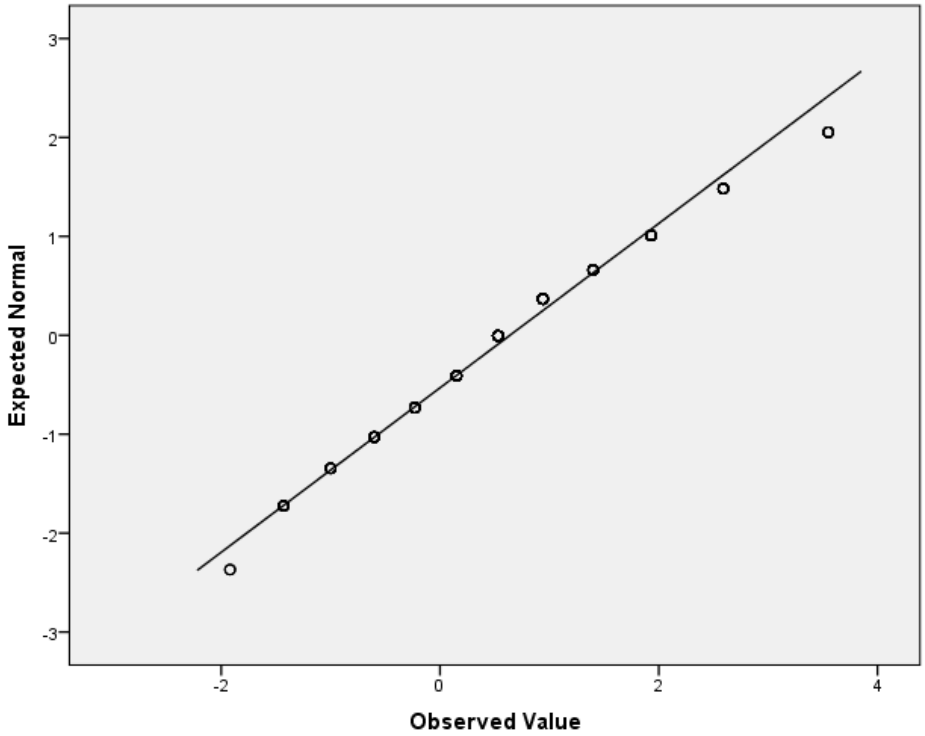

Figure 320: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Lecture Test 2 2014

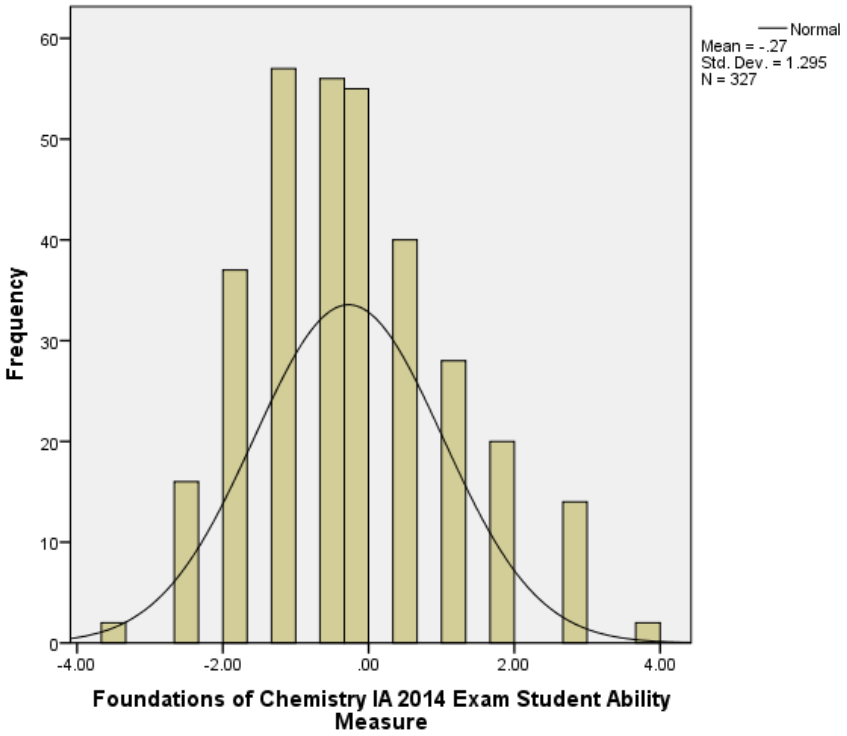

Figure 321: Histogram of the Rasch Student Ability Measures in Exam from Foundations of Chemistry IA 2014 to Determine the Distribution that the Measures Follow

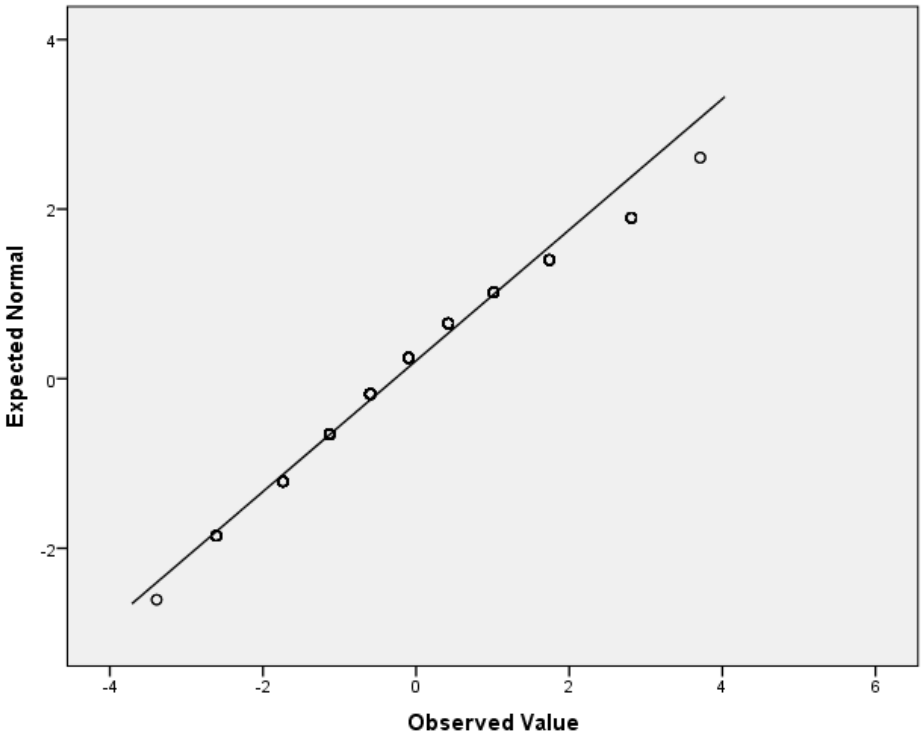

Figure 322: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Exam 2014

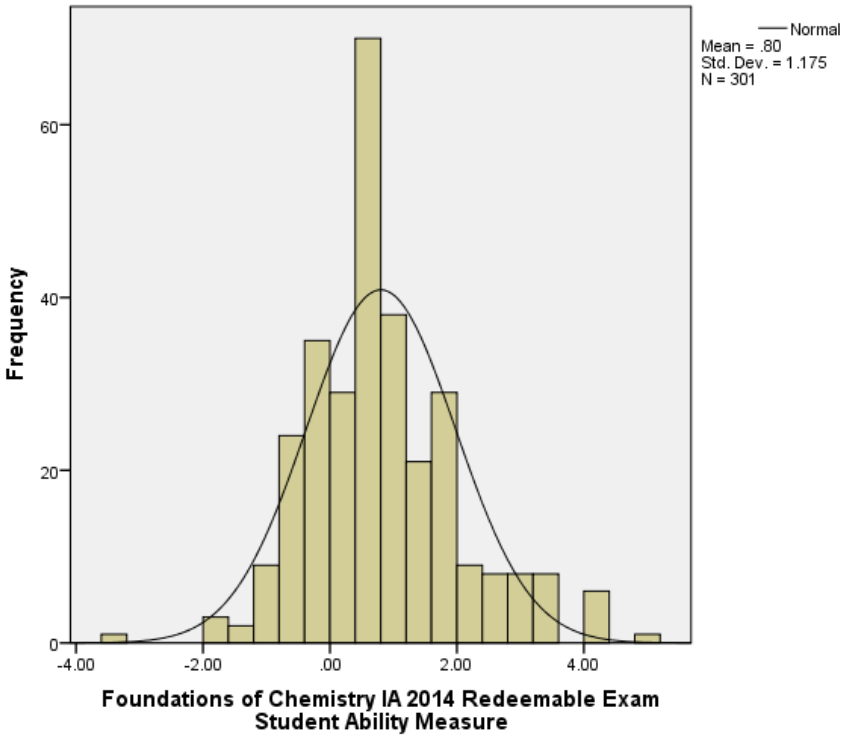

Figure 323: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Foundations of Chemistry IA 2014 to Determine the Distribution that the Measures Follow

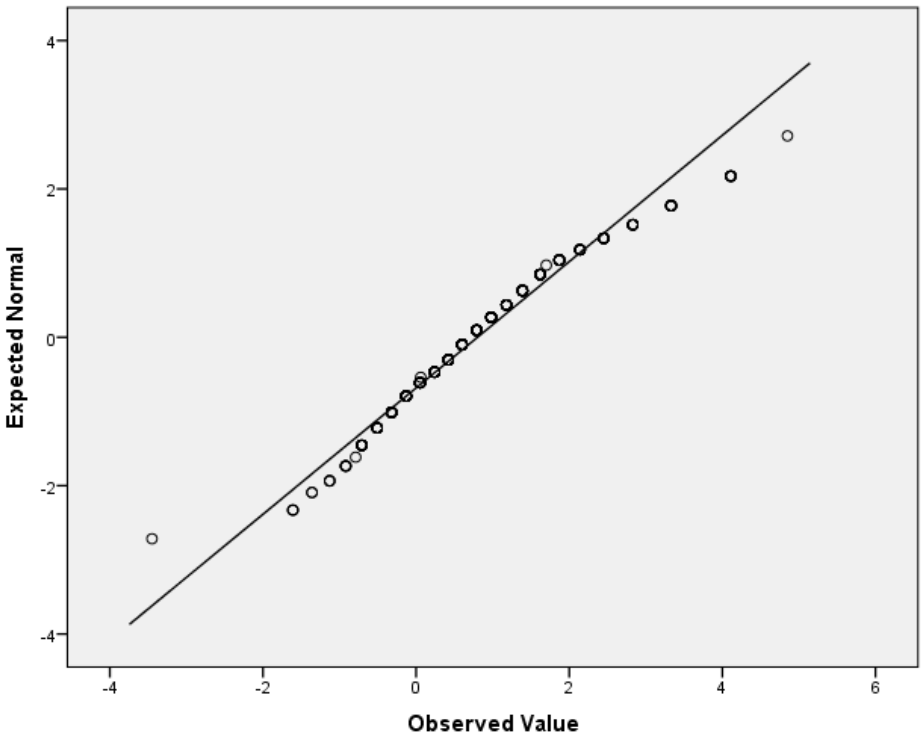

Figure 324: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Redeemable Exam 2014

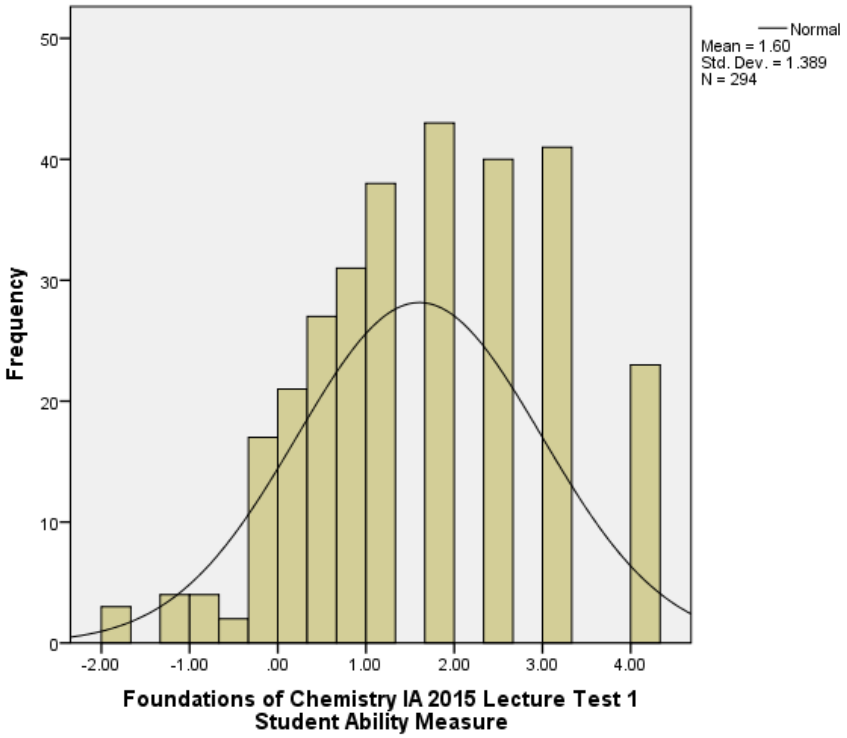

Figure 325: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Foundations of Chemistry IA 2015 to Determine the Distribution that the Measures Follow

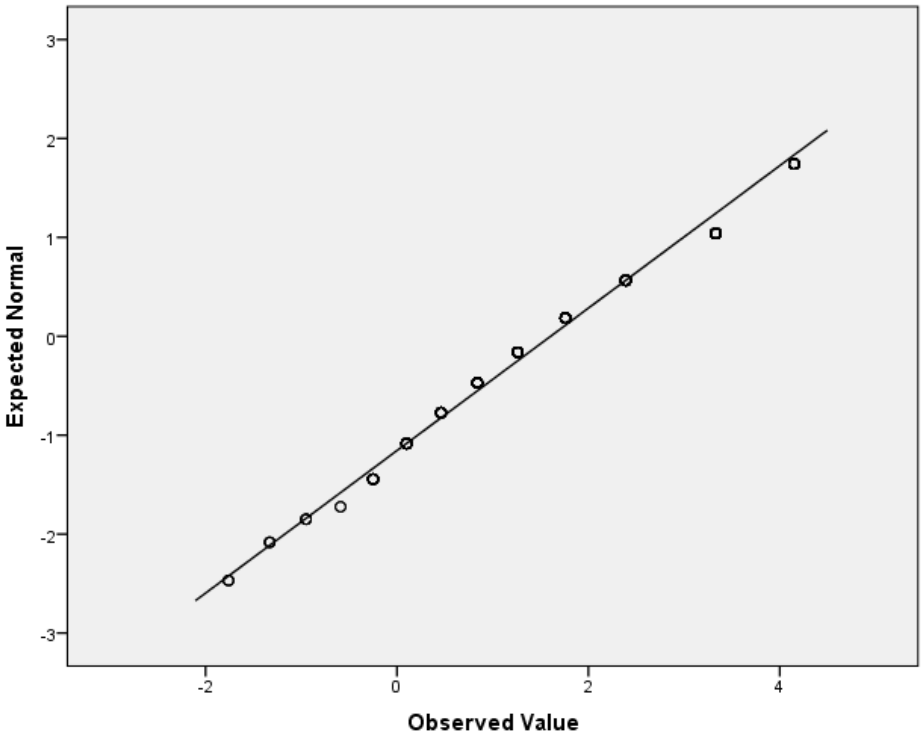

Figure 326: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Lecture Test 1 2015

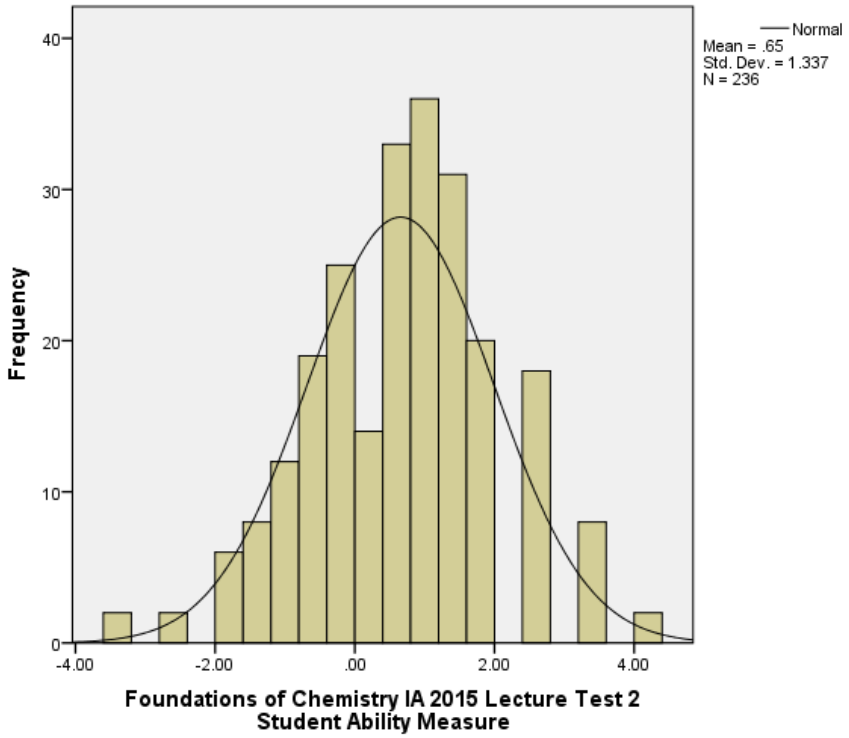

Figure 327: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Foundations of Chemistry IA 2015 to Determine the Distribution that the Measures Follow

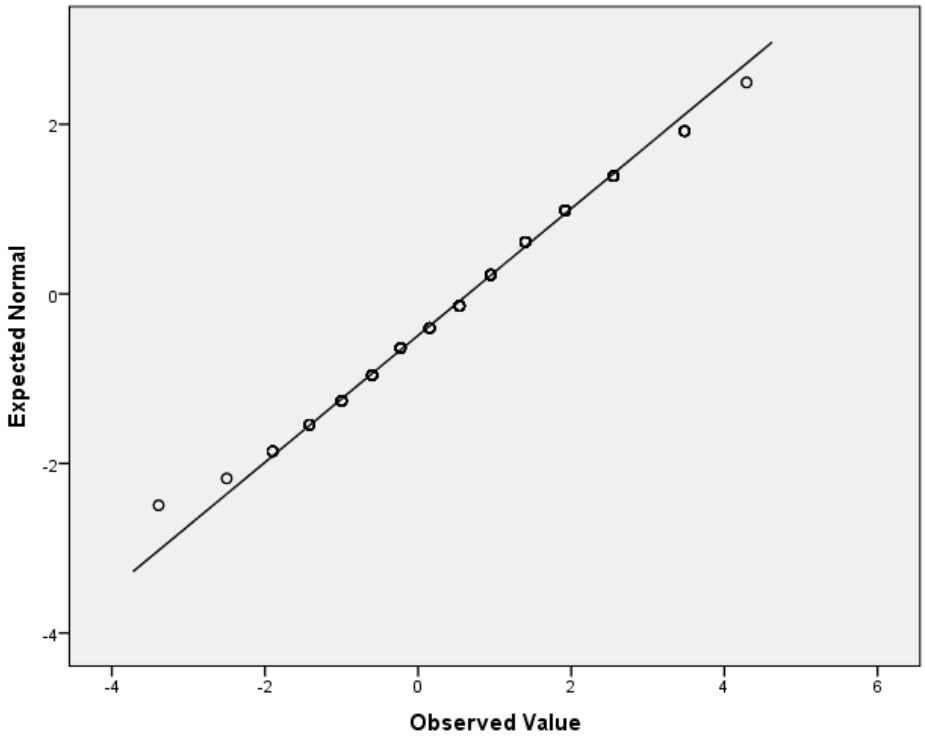

Figure 328: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Lecture Test 2 2015

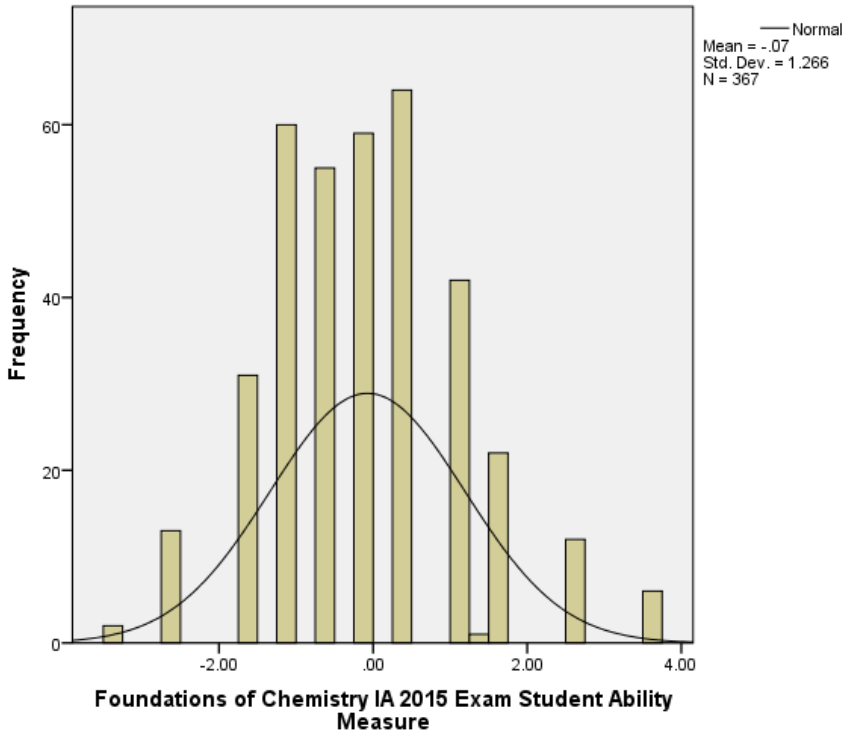

Figure 329: Histogram of the Rasch Student Ability Measures in Exam from Foundations of Chemistry IA 2015 to Determine the Distribution that the Measures Follow

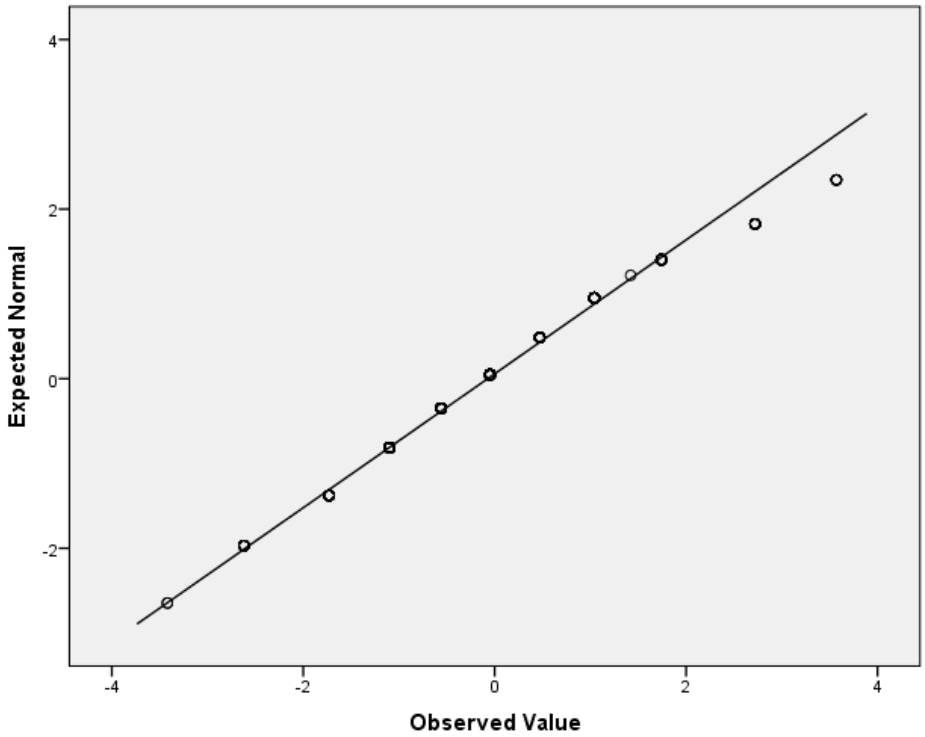

Figure 330: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Exam 2015

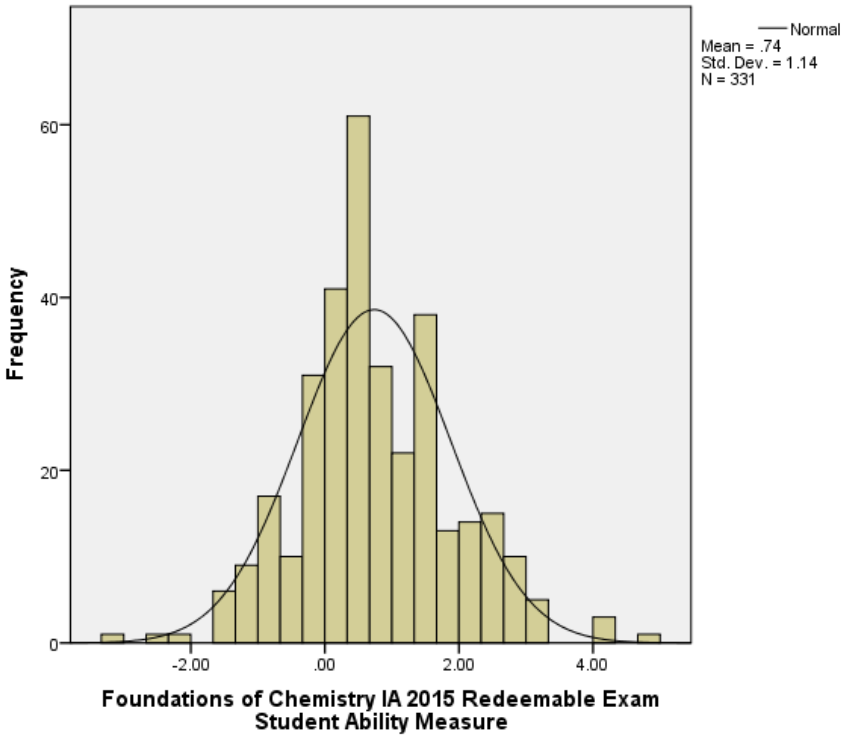

Figure 331: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Foundations of Chemistry IA 2015 to Determine the Distribution that the Measures Follow

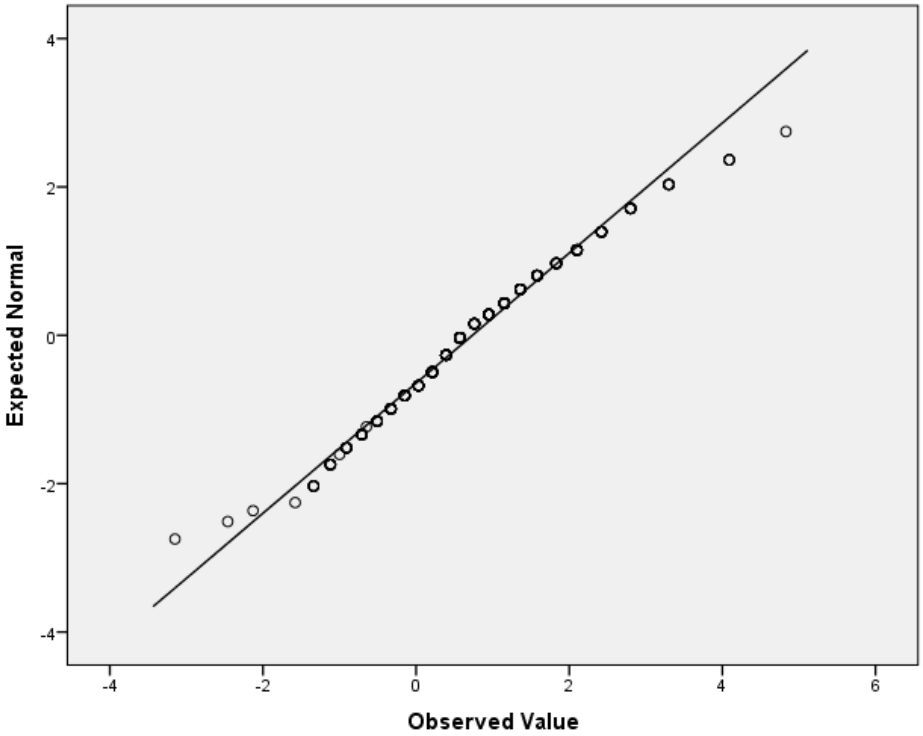

Figure 332: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IA Redeemable Exam 2015

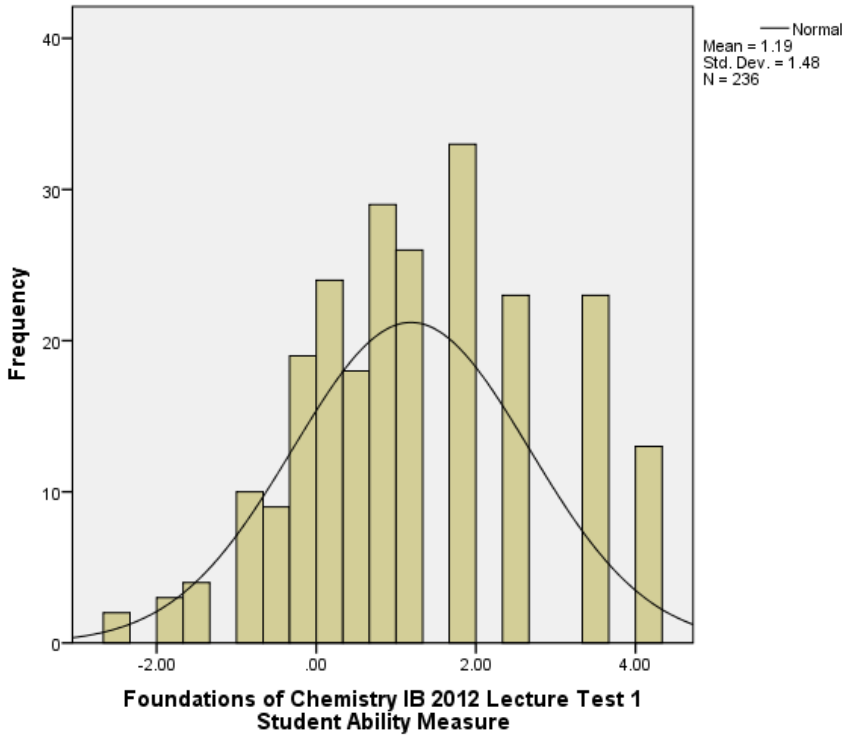

Figure 333: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Foundations of Chemistry IB 2012 to Determine the Distribution that the Measures Follow

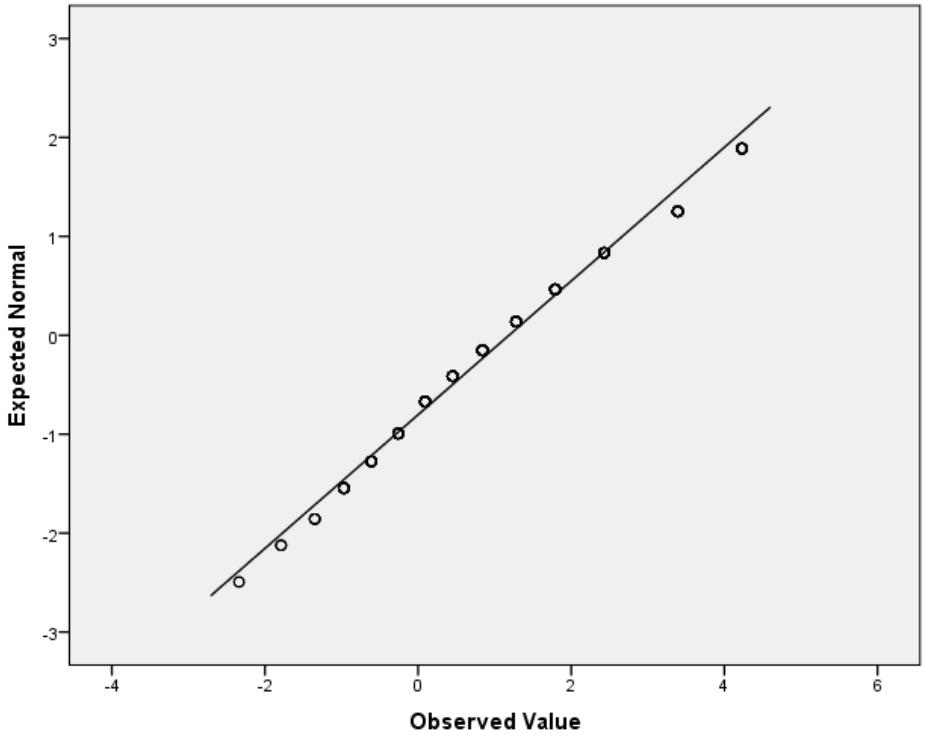

Figure 334: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Lecture Test 1 2012

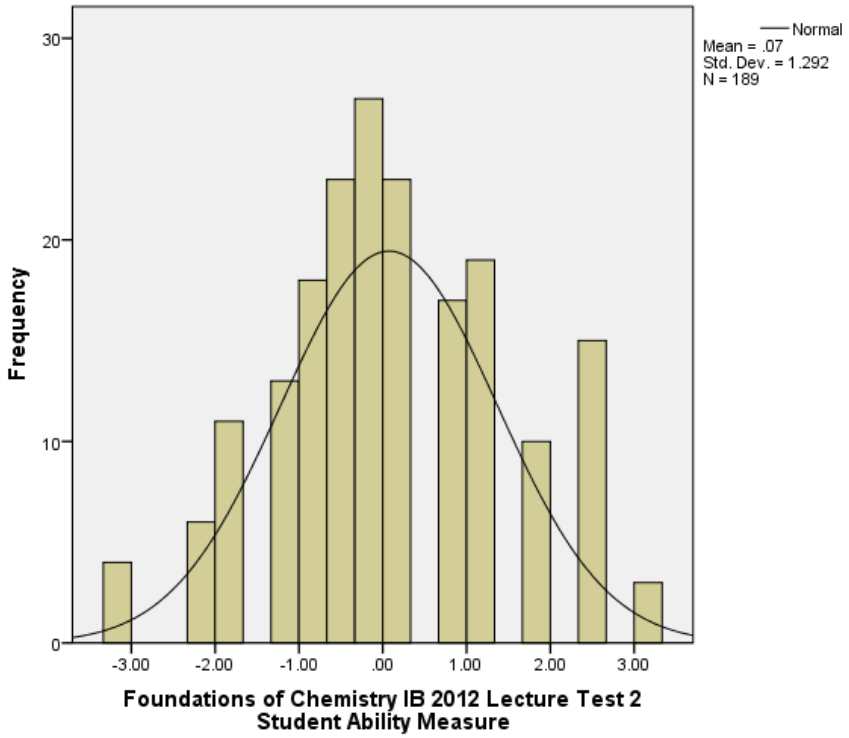

Figure 335: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Foundations of Chemistry IB 2012 to Determine the Distribution that the Measures Follow

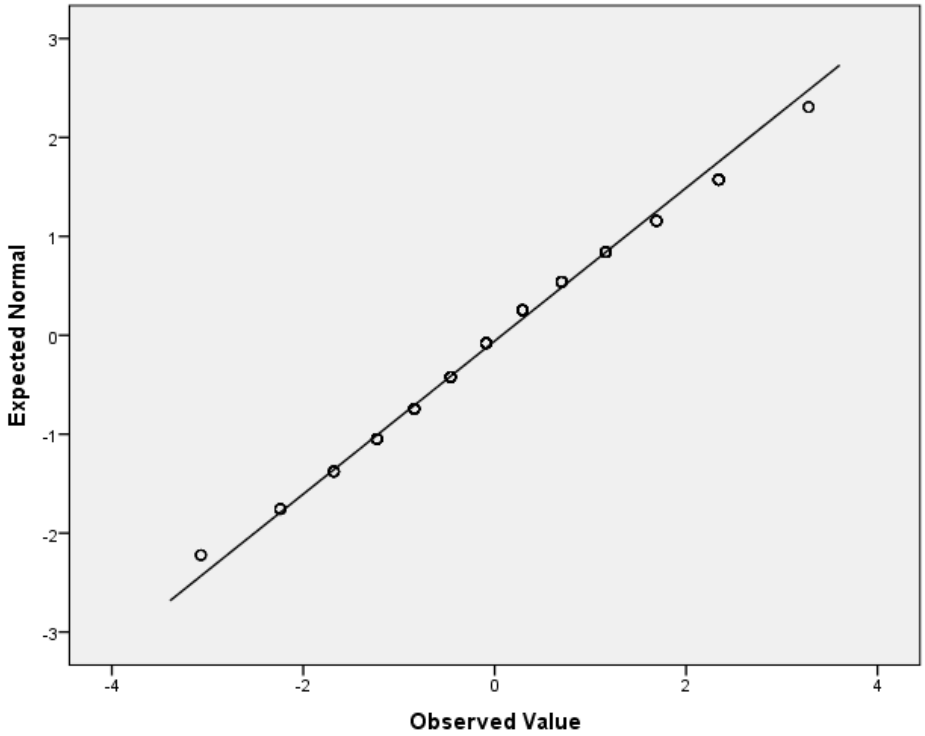

Figure 336: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Lecture Test 2 2012

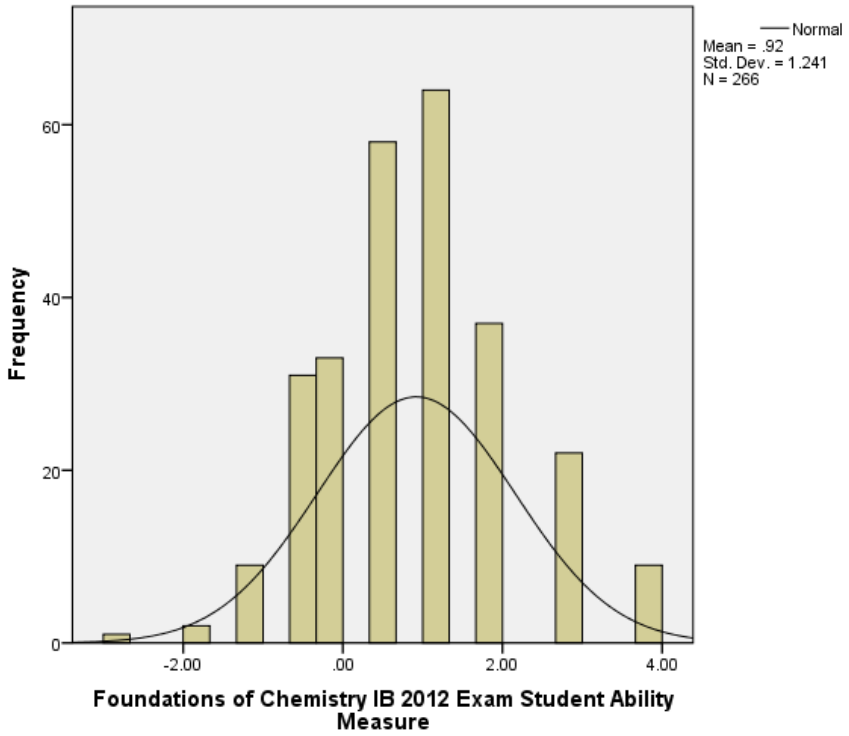

Figure 337: Histogram of the Rasch Student Ability Measures in Exam from Foundations of Chemistry IB 2012 to Determine the Distribution that the Measures Follow

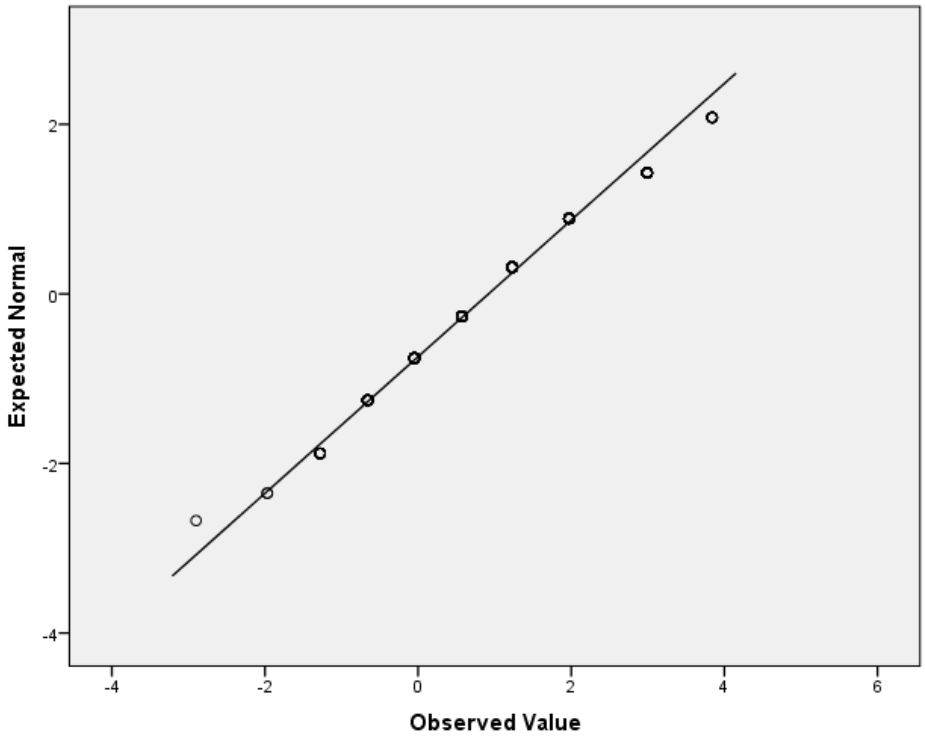

Figure 338: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Exam 2012

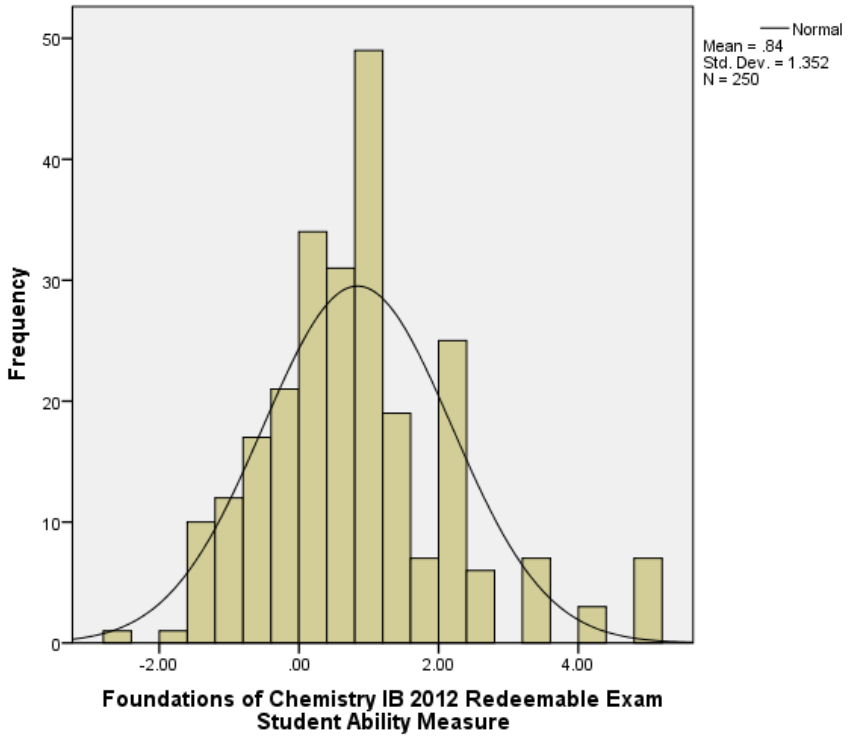

Figure 339: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Foundations of Chemistry IB 2012 to Determine the Distribution that the Measures Follow

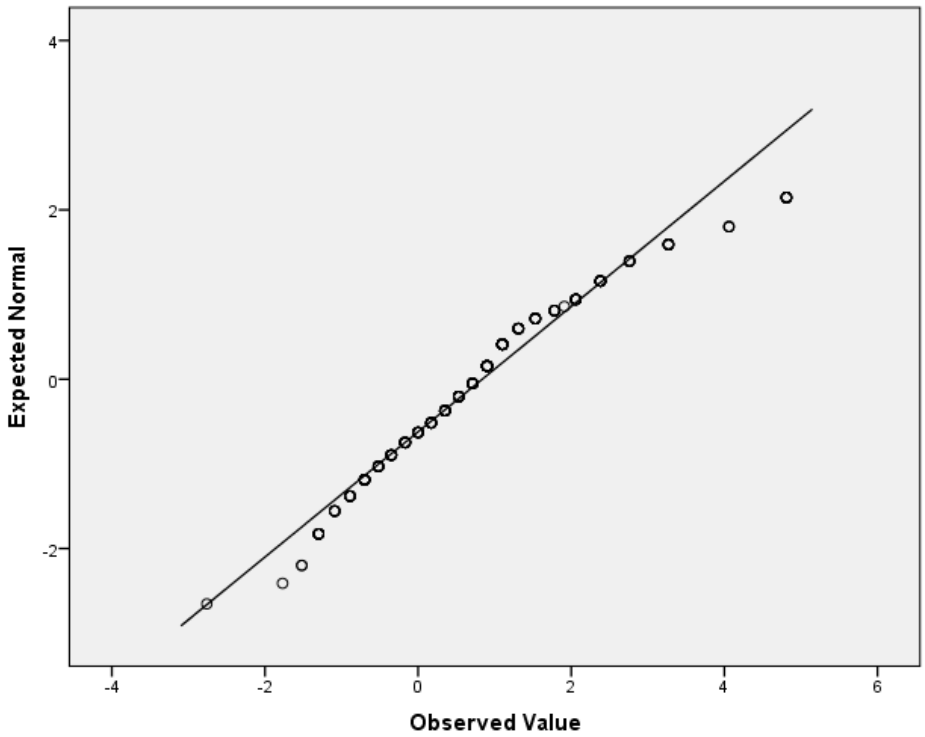

Figure 340: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Redeemable Exam 2012

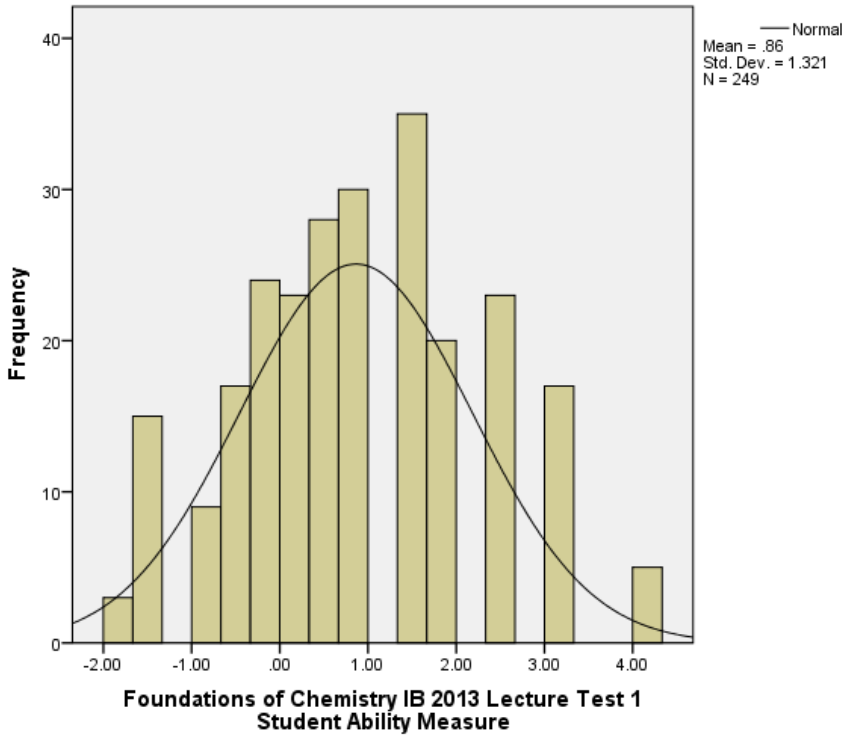

Figure 341: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Foundations of Chemistry IB 2013 to Determine the Distribution that the Measures Follow

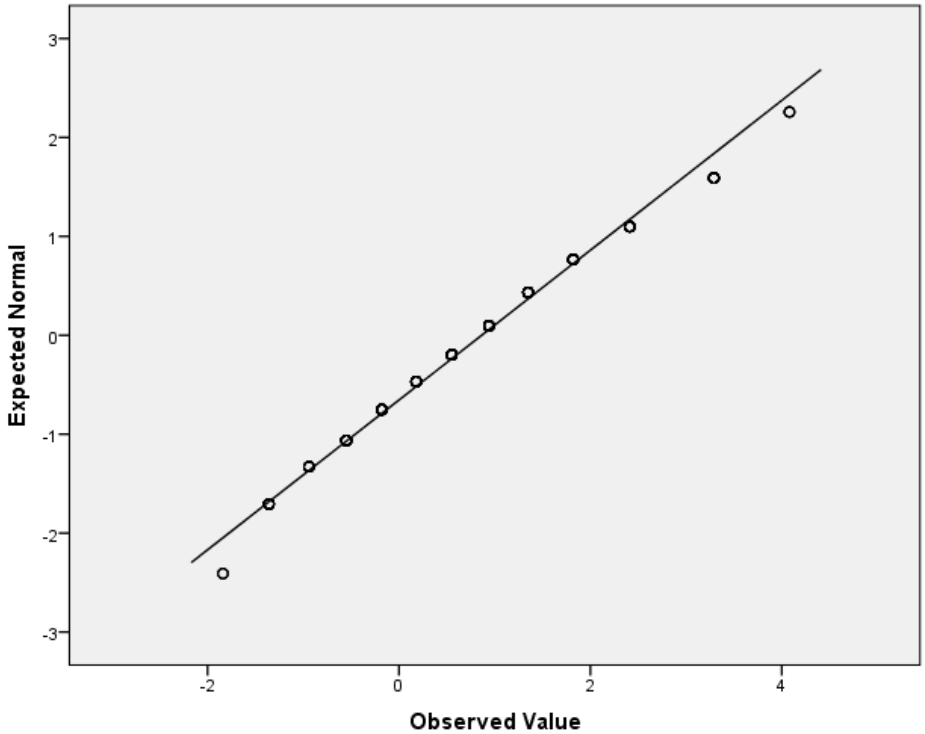

Figure 342: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Lecture Test 1 2013

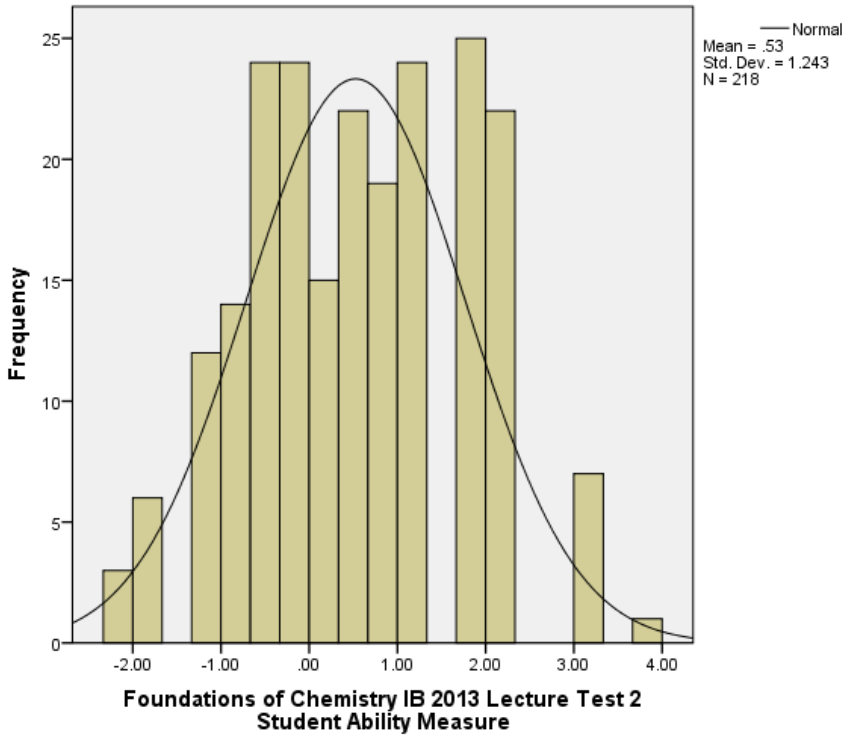

Figure 343: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Foundations of Chemistry IB 2013 to Determine the Distribution that the Measures Follow

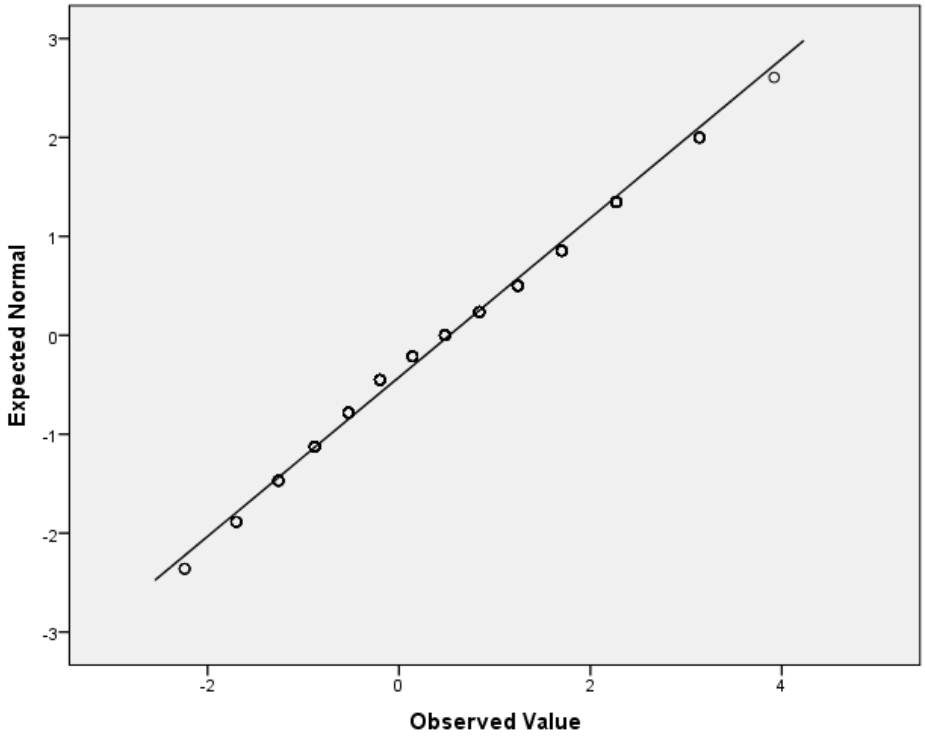

Figure 344: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Lecture Test 2 2013

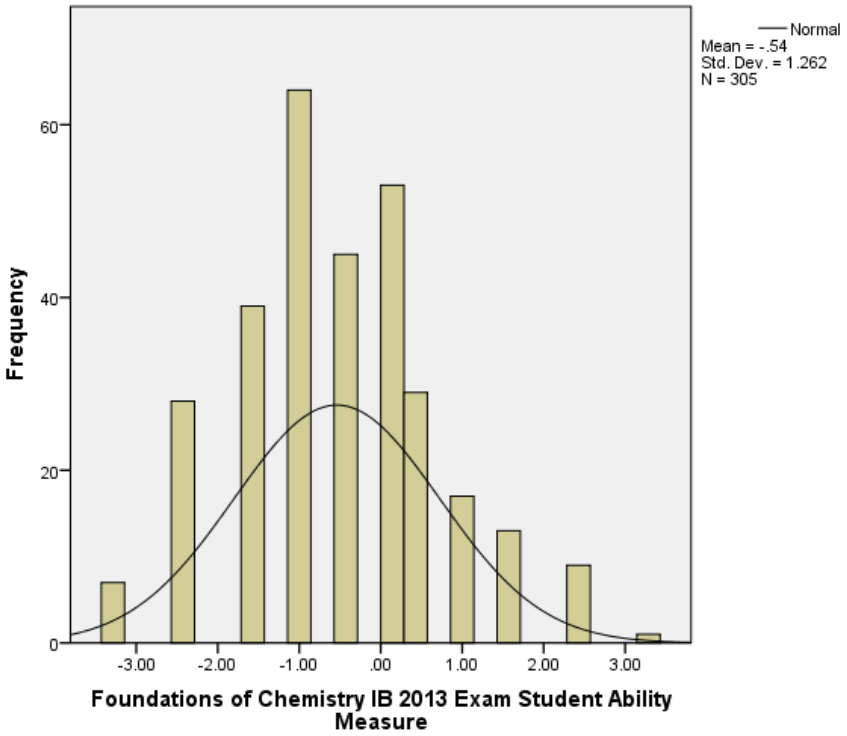

Figure 345: Histogram of the Rasch Student Ability Measures in Exam from Foundations of Chemistry IB 2013 to Determine the Distribution that the Measures Follow

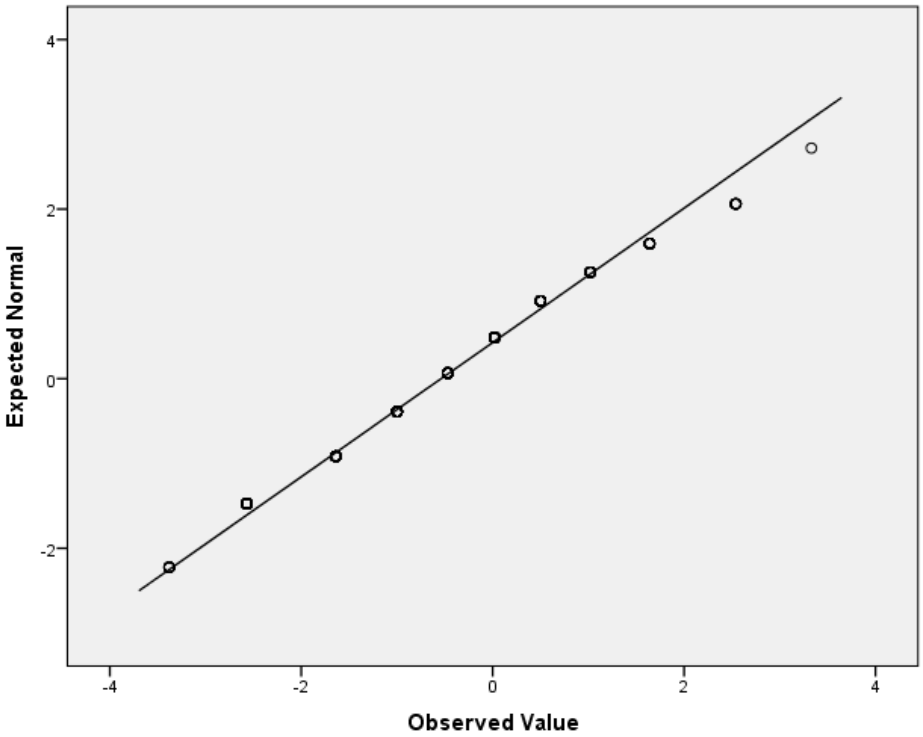

Figure 346: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Exam 2013

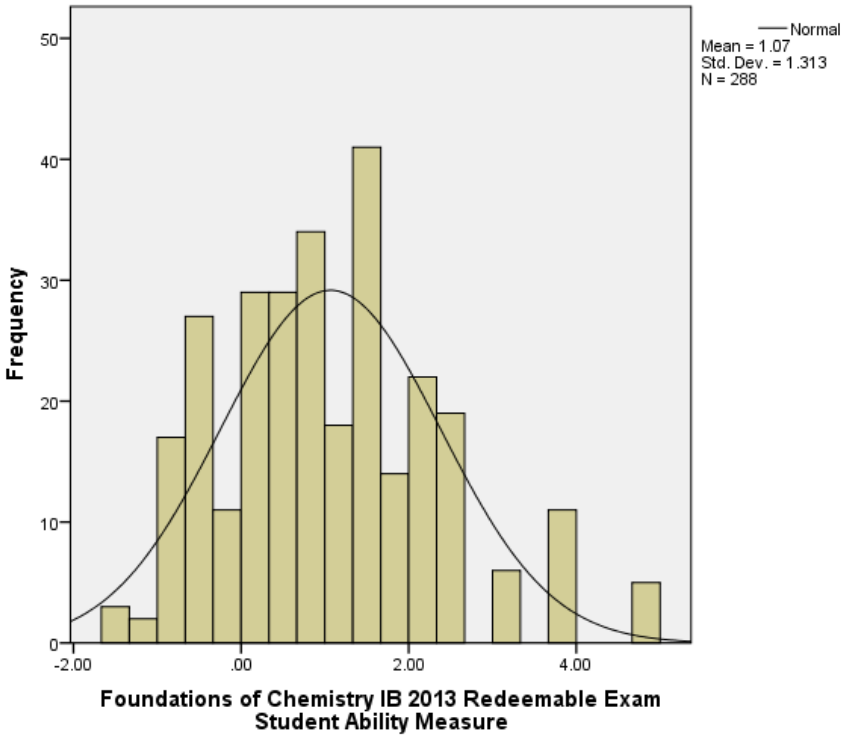

Figure 347: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Foundations of Chemistry IB 2013 to Determine the Distribution that the Measures Follow

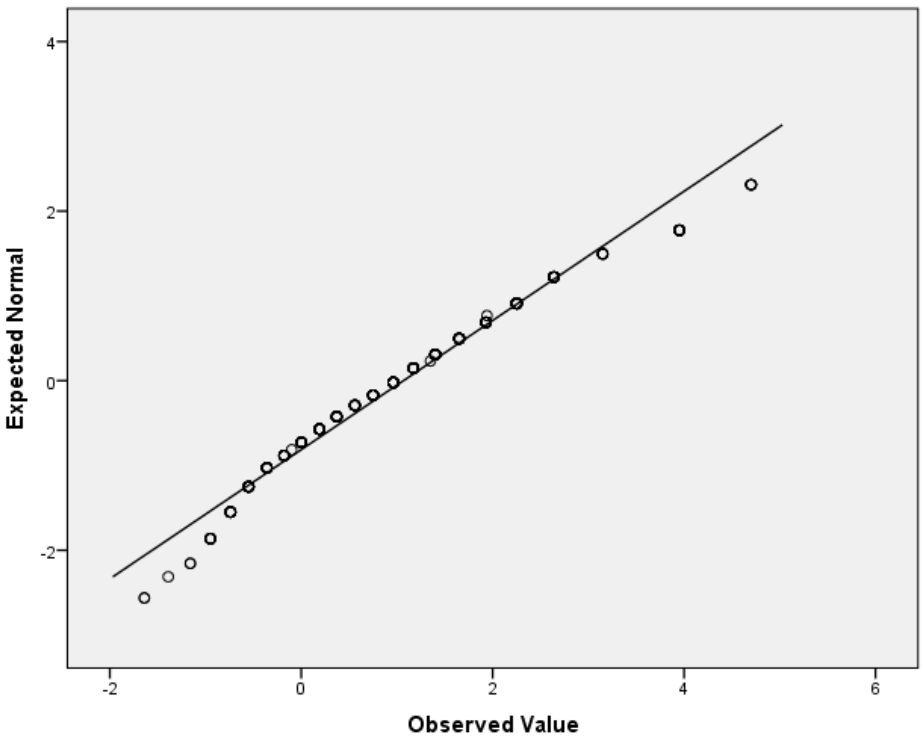

Figure 348: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Redeemable Exam 2013

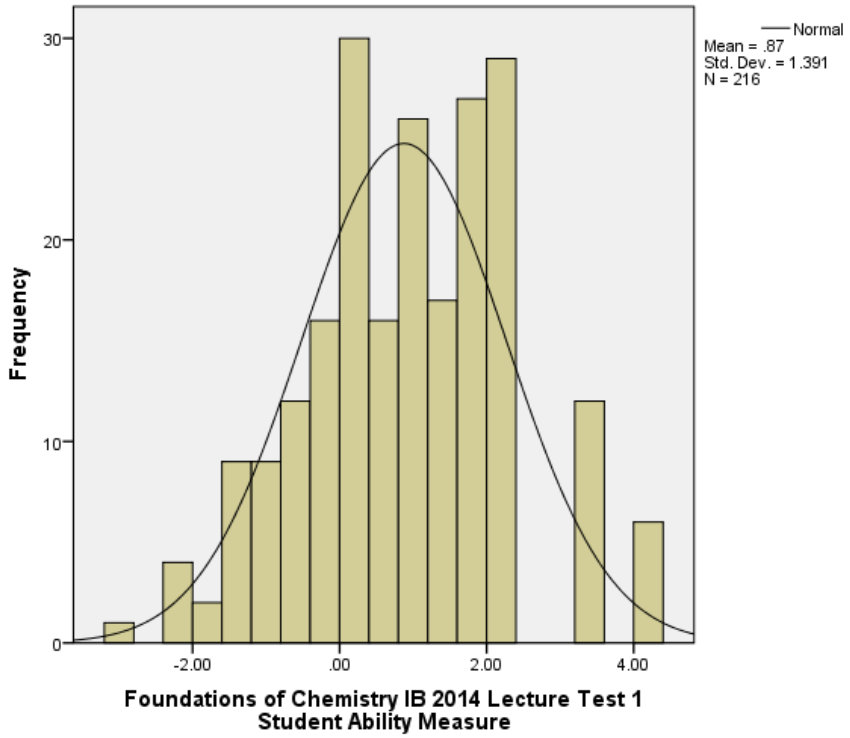

Figure 349: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Foundations of Chemistry IB 2014 to Determine the Distribution that the Measures Follow

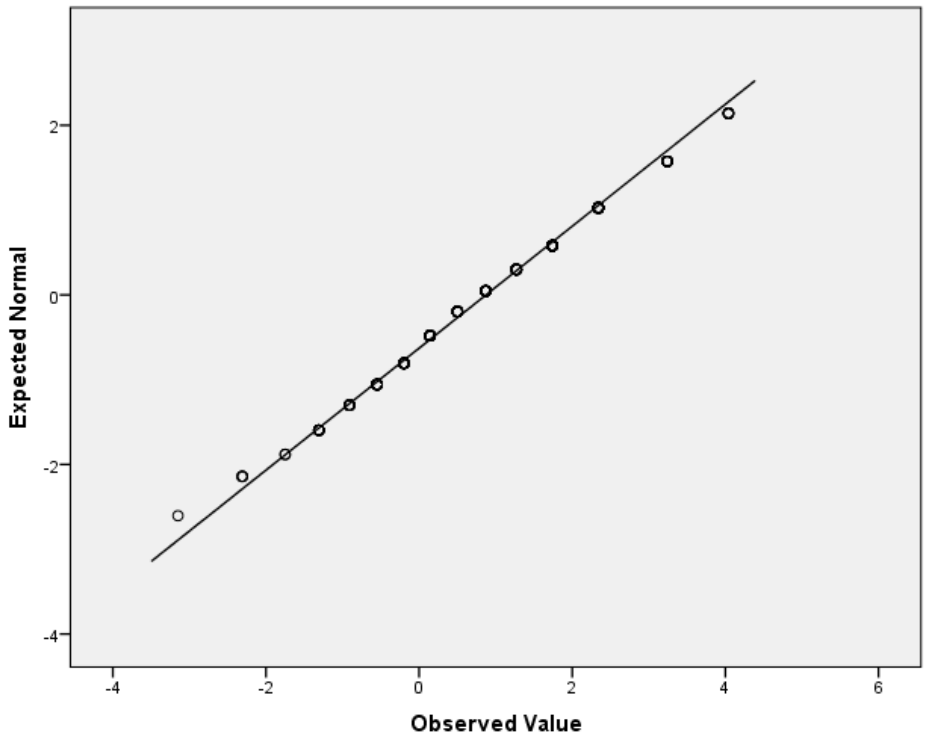

Figure 350: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Lecture Test 1 2014

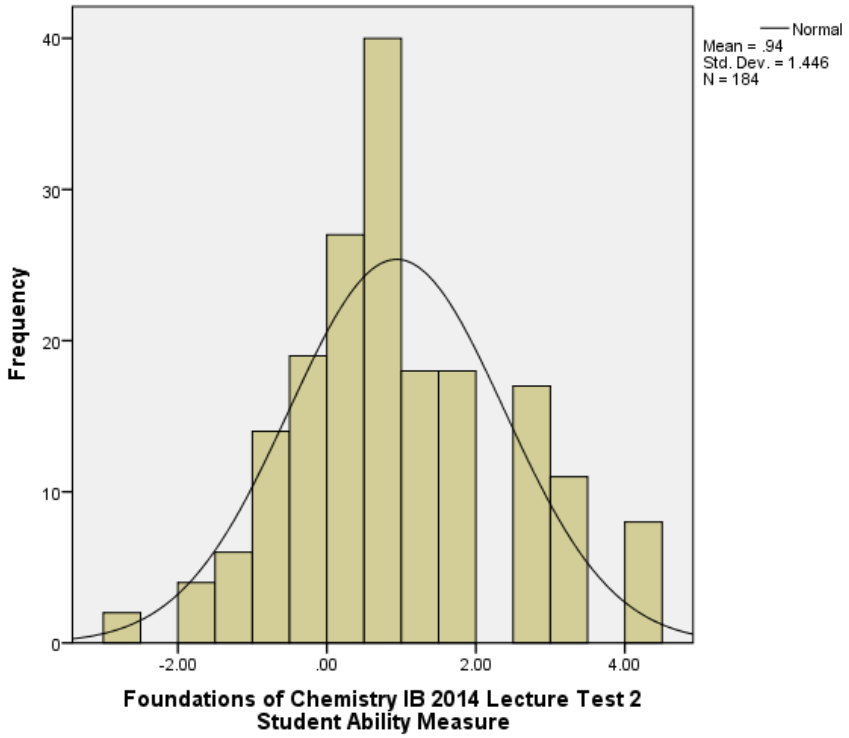

Figure 351: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Foundations of Chemistry IB 2014 to Determine the Distribution that the Measures Follow

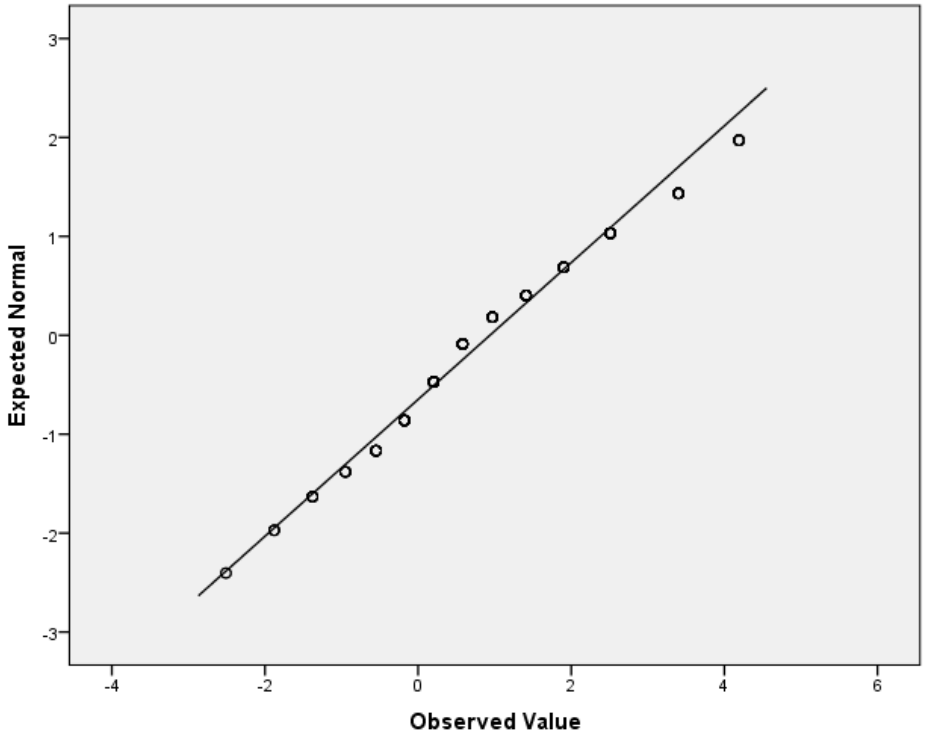

Figure 352: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Lecture Test 2 2014

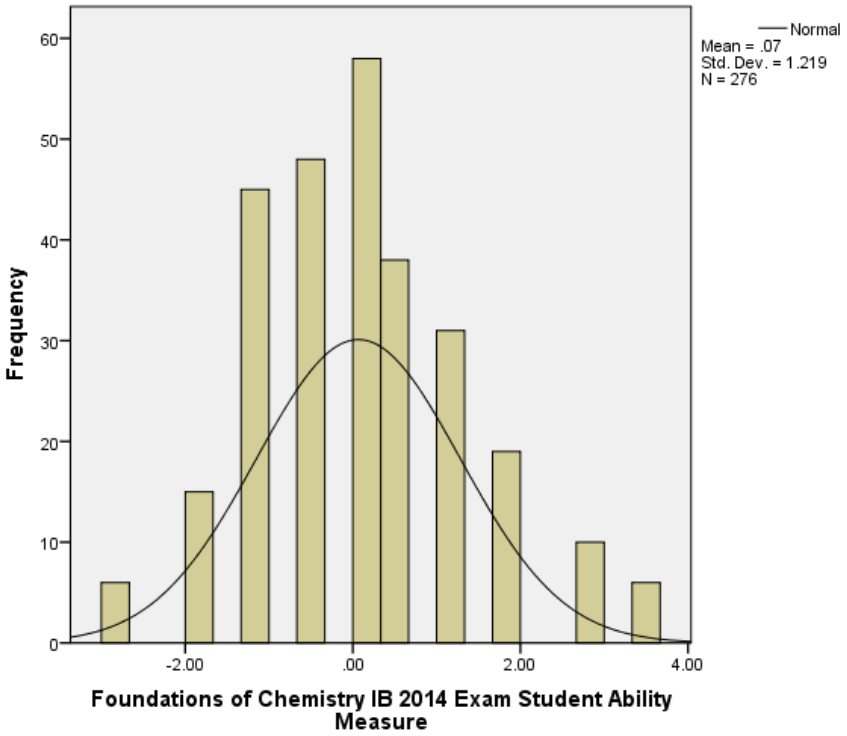

Figure 353: Histogram of the Rasch Student Ability Measures in Exam from Foundations of Chemistry IB 2014 to Determine the Distribution that the Measures Follow

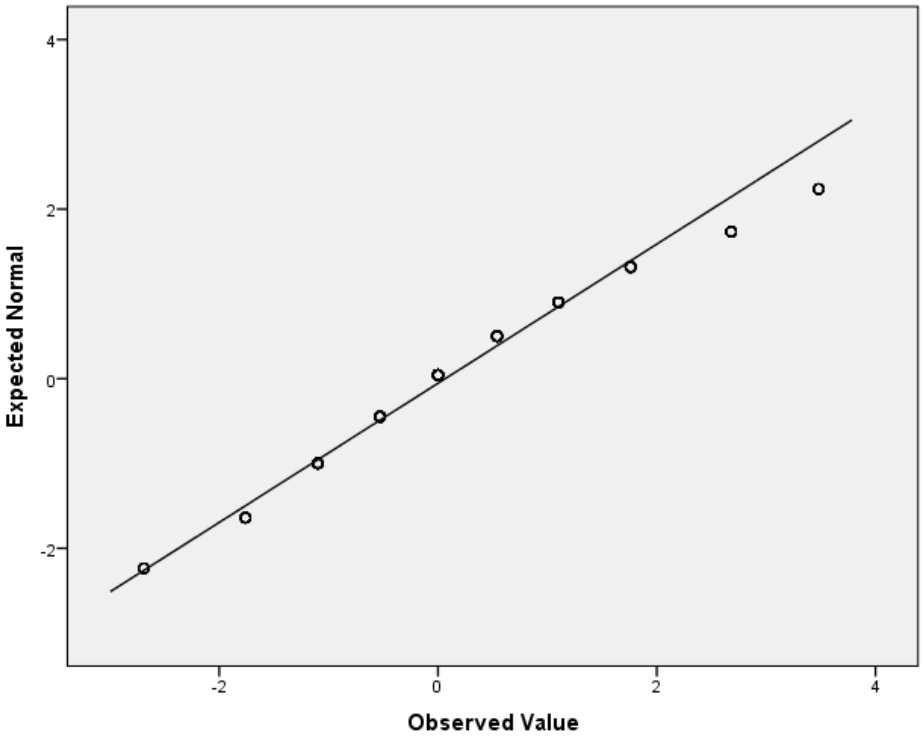

Figure 354: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Exam 2014

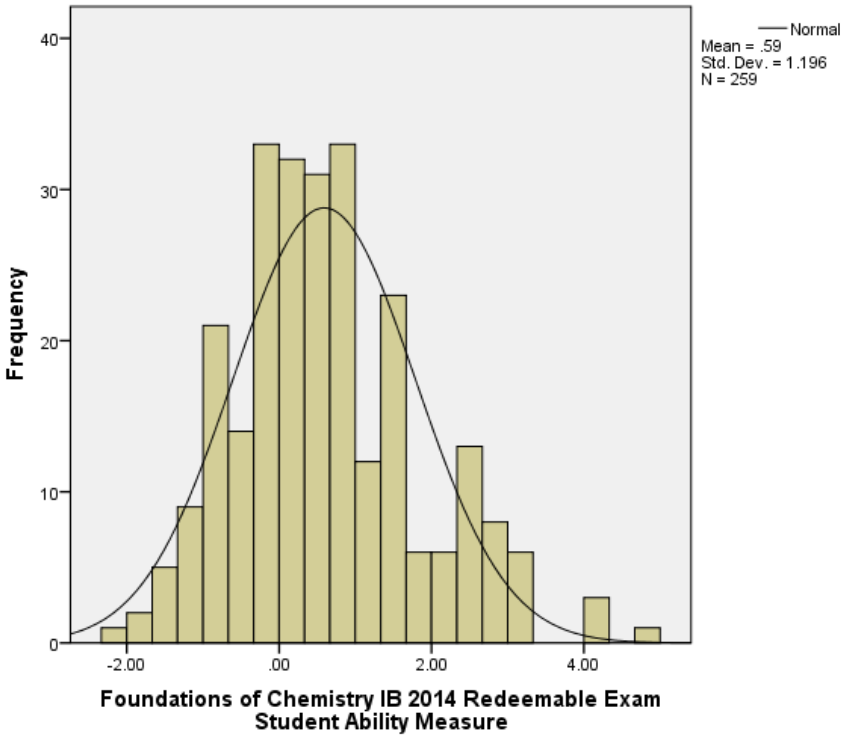

Figure 355: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Foundations of Chemistry IB 2014 to Determine the Distribution that the Measures Follow

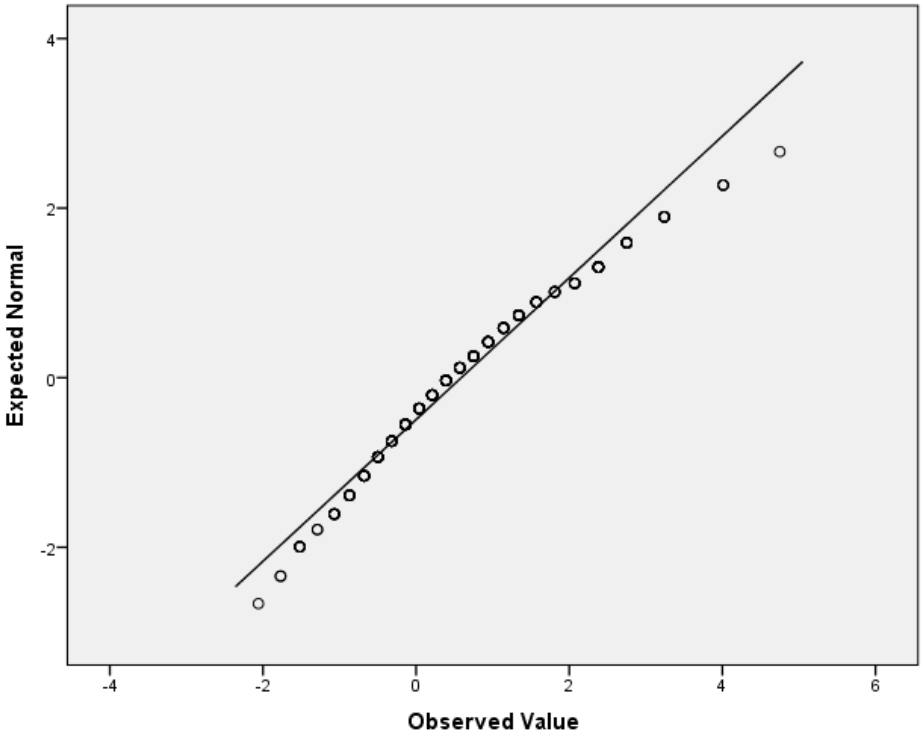

Figure 356: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Redeemable Exam 2014

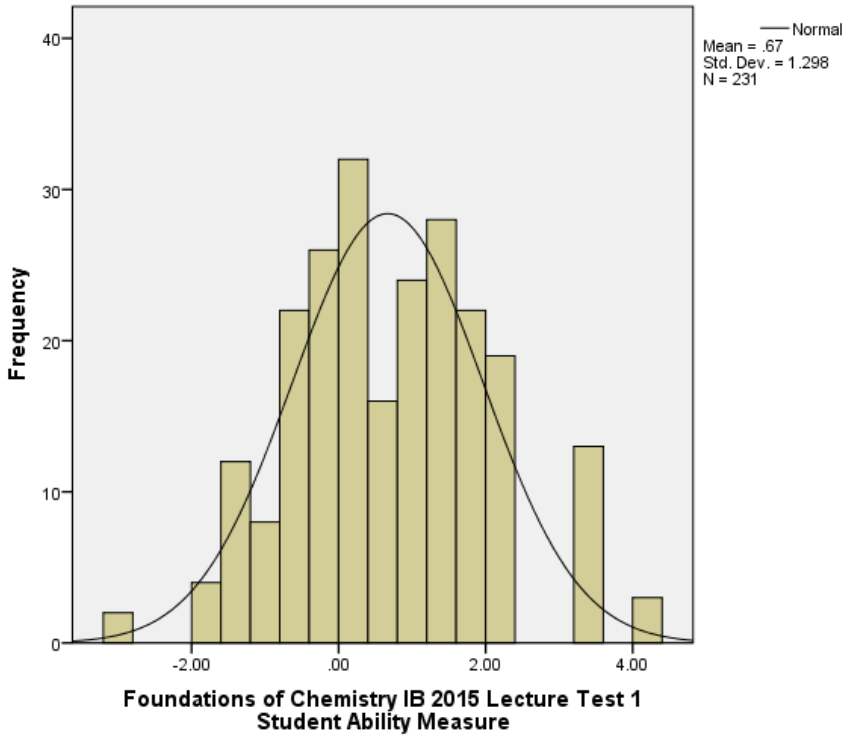

Figure 357: Histogram of the Rasch Student Ability Measures in Lecture Test 1 from Foundations of Chemistry IB 2015 to Determine the Distribution that the Measures Follow

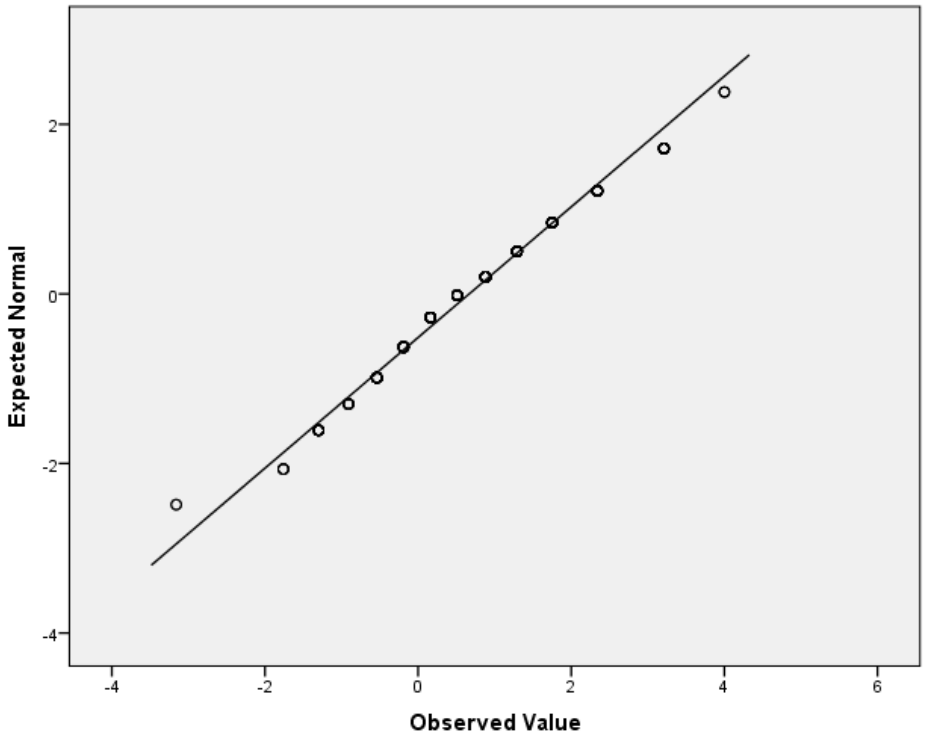

Figure 358: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Lecture Test 1 2015

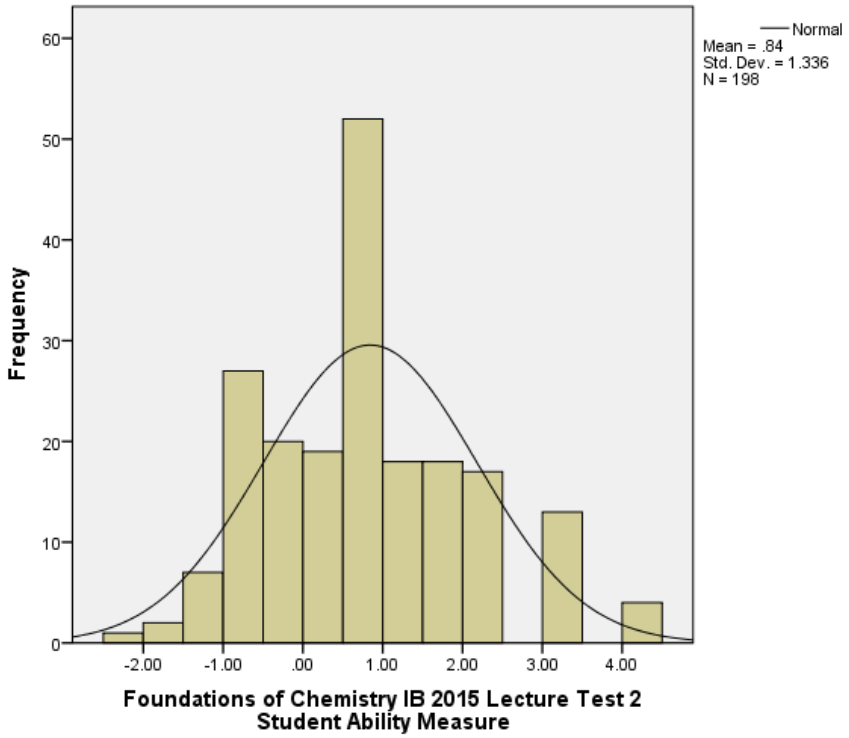

Figure 359: Histogram of the Rasch Student Ability Measures in Lecture Test 2 from Foundations of Chemistry IB 2015 to Determine the Distribution that the Measures Follow

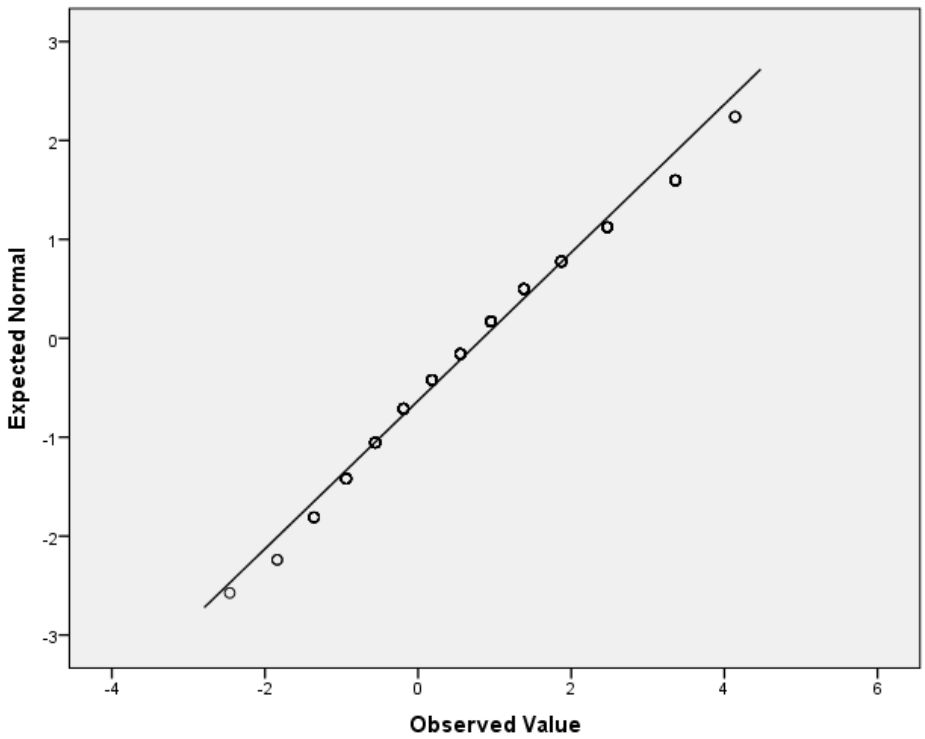

Figure 360: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Lecture Test 2 2015

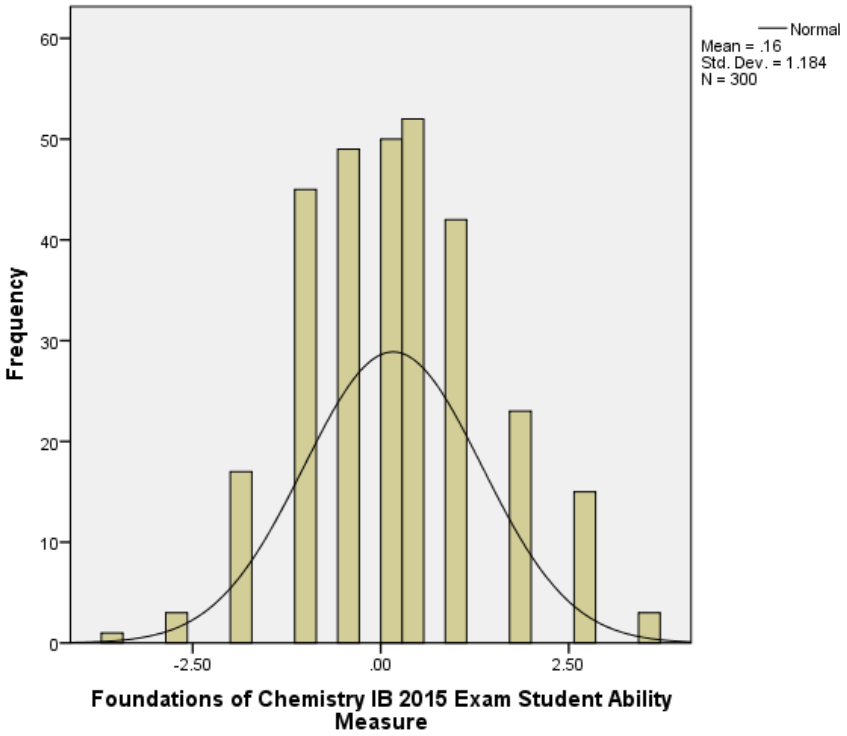

Figure 361: Histogram of the Rasch Student Ability Measures in Exam from Foundations of Chemistry IB 2015 to Determine the Distribution that the Measures Follow

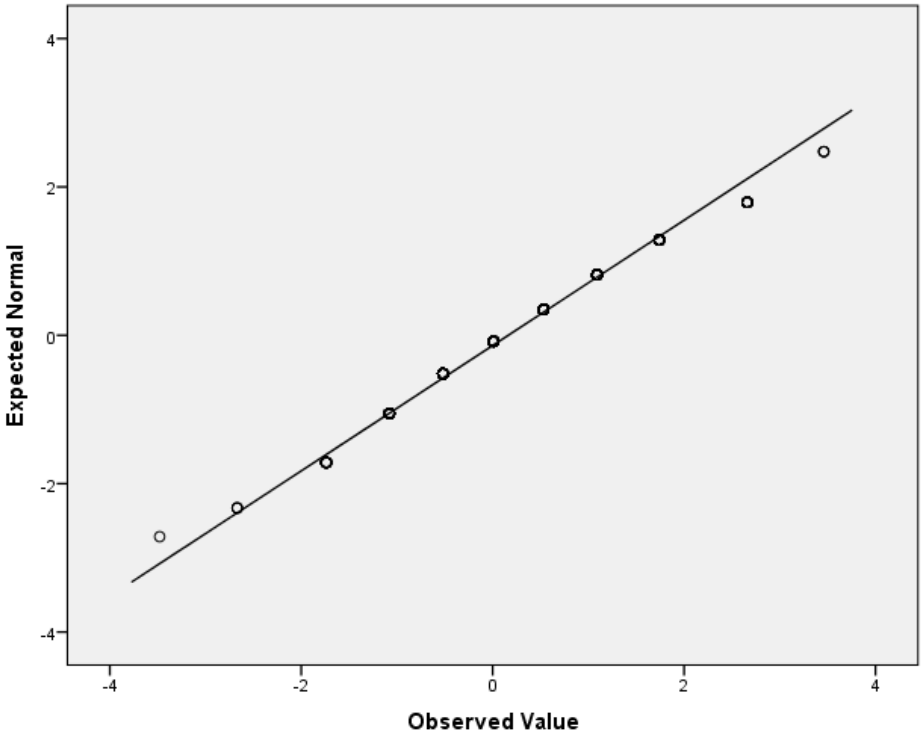

Figure 362: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Exam 2015

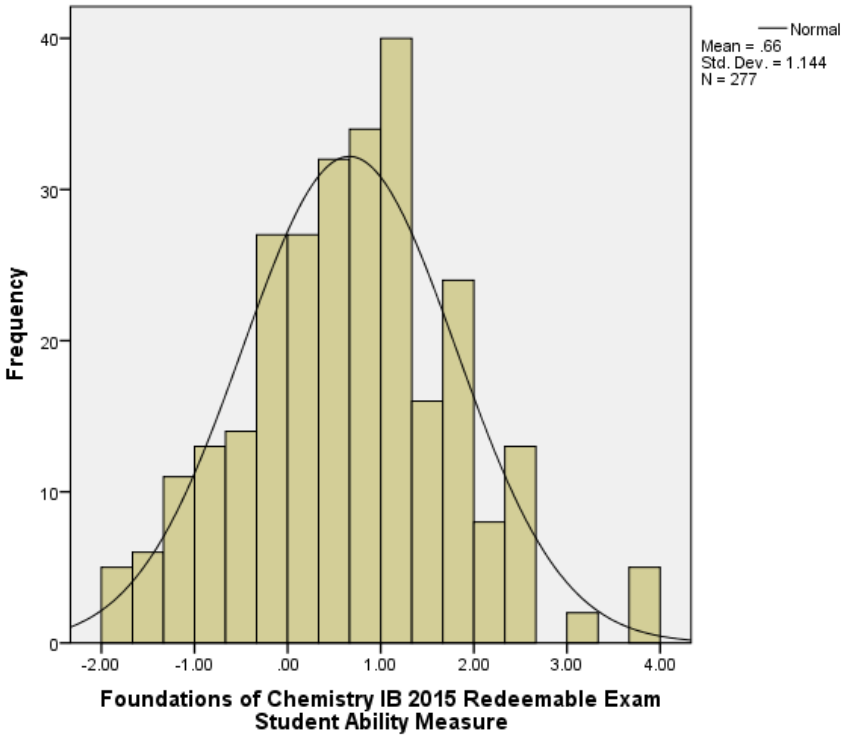

Figure 363: Histogram of the Rasch Student Ability Measures in Redeemable Exam from Foundations of Chemistry IB 2015 to Determine the Distribution that the Measures Follow

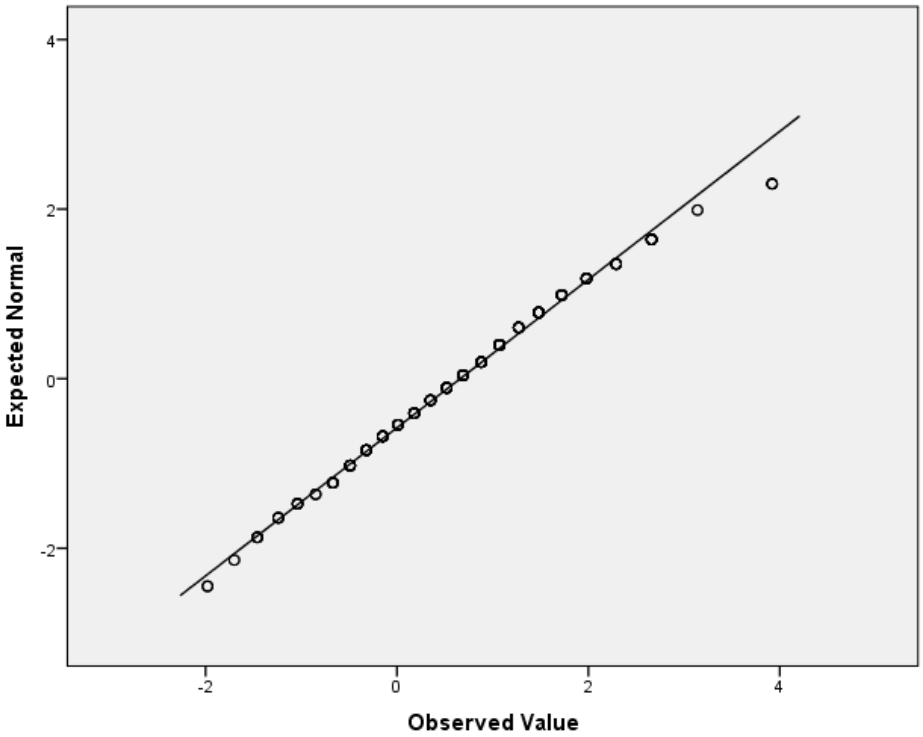

Figure 364: Rasch Student Ability Measure Q-Q Plot from Foundations of Chemistry IB Redeemable Exam 2015

7.8 Problematic Items Identified Using Classical Test Theory

Table 73: All of the Problematic Items Identified Using Classical Test Theory Analysis through the Evaluation of Item Difficulty, Discrimination, and Point Biserial Coefficient

	Item	Times Asked	Years Asked	Item	P	D	r_{pbi}
Chemistry IA	**Lec_2_5**	3	2012-2014				
			2012	Lec_2_5	0.327	0.152	0.114
			2013	Lec_2_5	0.266	-0.019	0.01
			2014	Lec_2_5	0.188	0.338	0.378
	Lec_2_8	8	2012-2015	*Redeemable Exam: Q23*			
			2012	Lec_2_8	0.224	0.395	0.307
			2012	Exam_2_23	0.25	0.279	0.285
	Lec_2_9	8	2012-2015	*Redeemable Exam: Q24* *Distractor Changes in 2014-2015*			
			2012	Lec_2_9	0.15	0.045	0.008
			2012	Exam_2_24	0.115	-0.041	-0.038
			2013	Lec_2_9	0.119	0.029	0.027
			2013	Exam_2_24	0.13	0.672	0.505
	Exam_2_18	1	2015				
			2015	Exam_2_18	0.141	0.183	0.188
	Exam_2_22	8	2012-2015	*Lecture Test 2: Q7*			
			2012	Exam_2_22	0.547	0.172	0.18
			2013	Exam_2_22	0.421	0.166	0.163
			2014	Exam_2_22	0.475	0.236	0.158
			2015	Exam_2_22	0.399	0.271	0.215
No Problematic Items from Chemistry IB							
Foundations of Chemistry IA	**Lec_1_1**	4	2012-2015				
			2012	Lec_1_1	0.942	-0.046	0.164
			2014	Lec_1_1	0.917	-0.0159	0.237
	Lec_2_13	4	2012-2015				
			2012	Lec_2_13	0.165	0.18	0.199
			2013	Lec_2_13	0.161	0.031	0.03
			2014	Lec_2_13	0.135	0.126	0.095
			2015	Lec_2_13	0.169	0.153	0.204
	Exam_1_10	4	2012-2015				
			2012	Exam_1_10	0.199	0.314	0.373
			2013	Exam_1_10	0.137	0.231	0.319
			2014	Exam_1_10	0.089	0.11	0.202
			2015	Exam_1_10	0.136	0.196	0.291
	Exam_2_18	4	2012-2015				
			2013	Exam_2_18	0.214	0.263	0.261
			2014	Exam_2_18	0.199	0.281	0.252
			2015	Exam_2_18	0.189	0.306	0.281

	Exam_2_29	4	2012-2015				
			2012	Exam_2_29	0.193	0.263	0.245
			2013	Exam_2_29	0.129	0.197	0.2
			2014	Exam_2_29	0.196	0.171	0.176
			2015	Exam_2_29	0.158	0.12	0.175
Foundations of Chemistry IB	**Exam_1_3**	3	2013-2015				
			2013	Exam_1_3	0.206	0.353	0.563
			2014	Exam_1_3	0.236	0.435	0.425
			2015	Exam_1_3	0.264	0.495	0.381
	Exam_1_4	3	2013-2015				
			2013	Exam_1_4	0.156	0.43	0.513
			2014	Exam_1_4	0.246	0.507	0.501
			2015	Exam_1_4	0.24	0.44	0.364

7.9 Problematic Items Identified Using Rasch Analysis

Table 74: The Items that Showed Misfit to the Rasch Model on More Than One Occasion Based on their Fit Measures on Each of Those Occasions

		Item	No. Times	Year	Item	Count	Score	Measure	S.E.	Discrim.	Infit	ZSTD	Outfit	ZSTD	Obs. Corr.	Exp. Corr.	Obs%	Exp %
Chemistry IA	Overfit	Lec_1_1	8	2012-2015	*Exam: Q1, Lec + Exam (2014-2015): Q10*													
				2014	Exam_2_10	509	380	-0.92	0.11	1.15	0.92	-1.40	0.79	-2.00	0.49	0.42	79.1	77.6
				2015	Exam_2_10	547	409	-1.05	0.11	1.17	0.90	-1.70	0.75	-2.60	0.54	0.47	79.2	78.8
		Lec_1_4	8	2012-2015	*Exam: Q4, Lec + Exam (2014-2015): Q13*													
				2012	Exam_2_4	488	335	-0.49	0.11	1.33	0.85	-3.40	0.76	-2.90	0.53	0.40	77.4	72.8
				2014	Exam_2_13	509	319	-0.22	0.10	1.35	0.87	-3.30	0.79	-3.10	0.53	0.43	75.1	70.2
		Lec_1_7	6	2012-2015	*Exam: Q9, Exam 1 (2014-2015): Q4*													
				2012	Exam_2_9	488	320	-0.32	0.11	1.34	0.86	-3.50	0.80	-2.70	0.53	0.41	77.8	71.4
				2014	Exam_1_4	509	331	-0.26	0.11	1.40	0.83	-4.20	0.77	-3.40	0.58	0.46	77.7	71.8
				2015	Exam_1_4	547	360	-0.38	0.10	1.39	0.83	-4.10	0.72	-4.30	0.59	0.46	76.7	72.0
		Lec_1_8	6	2012-2015	*Exam: Q6, Exam (2014-2015): Q15*													
				2012	Lec_1_8	470	336	-0.69	0.11	1.19	0.91	-1.70	0.78	-2.30	0.49	0.42	76.5	74.7
				2012	Exam_2_6	488	423	-1.76	0.14	1.07	0.95	-0.60	0.75	-1.40	0.37	0.30	86.2	86.9
				2013	Exam_2_6	506	421	-1.68	0.14	1.14	0.86	-1.60	0.73	-1.60	0.55	0.47	87.1	86.1
				2014	Exam_2_15	509	425	-1.59	0.13	1.09	0.93	-0.80	0.76	-1.60	0.47	0.41	85.9	85.5
		Lec_2_4	8	2012-2015	*Exam: Q21*													
				2012	Exam_2_21	488	291	-0.01	0.10	1.42	0.85	-4.00	0.78	-3.40	0.55	0.42	76.2	69.6
				2014	Exam_2_21	509	318	-0.21	0.10	1.45	0.83	-4.40	0.74	-3.90	0.55	0.43	76.5	70.1
		Lec_2_14	8	2012-2015	*Exam: Q29*													
				2013	Lec_2_14	421	338	-1.20	0.13	1.13	0.91	-1.20	0.79	-1.70	0.45	0.36	81.2	81.3
				2015	Lec_2_14	451	341	-0.67	0.12	1.24	0.85	-2.70	0.76	-2.30	0.48	0.36	81.3	76.9
				2015	Exam_2_29	547	379	-0.70	0.10	1.27	0.86	-2.90	0.78	-2.90	0.55	0.45	77.1	74.3
		Lec_2_15	4	2012-2015														
				2012	Lec_2_15	446	329	-0.91	0.12	1.30	0.83	-3.10	0.70	-3.60	0.54	0.37	79.4	76.0
				2013	Lec_2_15	421	313	-0.80	0.12	1.34	0.80	-3.60	0.64	-4.00	0.58	0.39	80.8	76.8
				2014	Lec_2_15	436	339	-0.96	0.13	1.27	0.80	-3.10	0.66	-3.30	0.56	0.38	82.5	79.7
				2015	Lec_2_15	451	364	-1.02	0.13	1.21	0.85	-2.20	0.68	-2.60	0.46	0.33	83.5	81.1
		Exam_1_2	2	2014-2015														
				2014	Exam_1_2	509	372	-0.74	0.11	1.22	0.88	-2.30	0.78	-2.50	0.51	0.43	79.8	75.7
				2015	Exam_1_2	547	383	-0.63	0.11	1.27	0.86	-3.10	0.80	-2.70	0.55	0.45	78.5	74.3
	Underfit	Lec_1_2	8	2012-2015	*Exam: Q2, Lec + Exam (2014-2015): Q11*													
				2013	Exam_2_2	506	387	-1.12	0.12	0.86	1.09	1.30	1.22	1.60	0.40	0.46	79.2	80.1
				2014	Exam_2_11	509	376	-0.87	0.11	0.81	1.08	1.40	1.43	3.60	0.34	0.42	77.1	77.0
		Lec_1_5	8	2012-2015	*Exam: Q5, Lec + Exam (2014-2015): Q14*													
				2012	Lec_1_5	470	374	-1.23	0.13	0.90	1.05	0.70	1.29	2.00	0.32	0.38	80.3	80.5
				2013	Exam_2_5	506	398	-1.29	0.12	0.86	1.07	1.00	1.32	2.00	0.40	0.46	82.7	82.0
				2014	Lec_1_14	474	300	-0.39	0.11	0.68	1.11	2.60	1.28	3.30	0.32	0.42	66.8	70.9
				2014	Exam_2_14	509	411	-1.36	0.13	0.94	1.03	0.40	1.20	1.40	0.38	0.41	82.9	82.9
				2015	Lec_1_14	504	348	-0.56	0.11	0.78	1.08	1.80	1.31	3.40	0.32	0.40	70.9	73.1
				2015	Exam_2_14	547	420	-1.19	0.12	0.89	1.05	0.70	1.44	3.50	0.42	0.47	81.7	80.7
		Lec_2_5	3	2012-2014														
				2012	Lec_2_5	446	146	1.17	0.11	0.40	1.25	5.00	1.42	5.10	0.12	0.38	66.6	71.7
				2013	Lec_2_5	421	112	1.71	0.12	0.34	1.33	5.30	1.84	6.70	0.03	0.38	73.4	75.1
				2014	Lec_2_5	436	118	1.74	0.12	0.45	1.27	4.50	1.61	5.00	0.09	0.37	71.0	75.4
		Lec_2_6	8	2012-2015	*Exam: Q19*													
				2012	Exam_2_19	488	254	0.37	0.10	0.43	1.19	4.80	1.25	4.00	0.28	0.44	58.7	68.7
				2013	Exam_2_19	506	254	0.39	0.10	0.43	1.20	4.90	1.24	3.70	0.33	0.45	59.6	68.7
				2015	Lec_2_6	451	284	0.04	0.11	0.52	1.17	3.90	1.26	3.40	0.27	0.41	62.8	69.4
		Lec_2_7	8	2012-2015	*Exam: Q22*													
				2012	Lec_2_7	446	272	-0.22	0.11	0.53	1.17	3.90	1.20	3.20	0.24	0.40	60.1	68.9
				2012	Exam_2_22	488	267	0.24	0.10	0.10	1.28	6.80	1.54	7.60	0.18	0.43	58.1	68.8
				2013	Lec_2_7	421	281	-0.36	0.11	0.59	1.19	3.70	1.23	2.70	0.24	0.41	65.3	72.2
				2013	Exam_2_22	506	213	0.81	0.10	-0.02	1.37	8.20	1.58	8.10	0.19	0.44	56.1	69.8
				2014	Lec_2_7	436	284	-0.21	0.11	0.69	1.13	2.60	1.23	3.00	0.29	0.41	68.7	71.6
				2014	Exam_2_22	509	242	0.54	0.10	-0.04	1.33	8.20	1.41	6.70	0.18	0.42	52.2	67.9
				2015	Lec_2_7	451	277	0.12	0.11	0.55	1.15	3.50	1.27	3.60	0.29	0.42	62.3	68.8
				2015	Exam_2_22	547	218	0.83	0.10	0.28	1.23	5.90	1.33	5.70	0.23	0.39	59.4	68.2
		Lec_2_8	8	2012-2015	*Exam: Q23*													
				2012	Lec_2_8	446	100	1.83	0.1	0.92	1	0.4	1.26	2.5	0.28	0.34	78.7	78.8
				2012	Exam_2_23	488	122	1.85	0.12	0.81	1.11	1.80	1.33	3.00	0.31	0.42	77.4	78.4
		Lec_2_9	8	2012-2015	*Exam: Q24 (Distractor Changes in 2014 Removed Underfit from Item)*													
				2012	Lec_2_9	446	67	2.34	0.14	0.70	1.23	2.50	1.81	4.40	0.01	0.30	84.8	85.0
				2012	Exam_2_24	488	56	2.98	0.15	0.62	1.34	3.00	2.70	6.20	-0.02	0.34	86.4	88.9
				2013	Lec_2_9	421	50	2.84	0.16	0.75	1.21	1.90	2.03	4.10	0.02	0.29	86.9	88.4
				2013	Exam_2_24	506	66	2.77	0.14	0.62	1.32	3.10	2.41	5.90	0.06	0.33	86.3	87.3
		Exam_1_1	4	2012-2015	*Exam 1 (2014-2015): Q5*													
				2013	Exam_1_1	506	368	-1.35	0.11	0.78	1.12	2.20	1.24	2.50	0.30	0.41	72.0	76.1
				2014	Exam_1_5	509	327	-0.21	0.11	0.39	1.26	5.60	1.38	4.70	0.28	0.46	62.1	71.5
				2015	Exam_1_5	547	341	-0.18	0.10	0.60	1.15	3.60	1.27	3.90	0.35	0.47	65.0	70.7
		Exam_2_18	1	2015														
				2015	Exam_2_18	547	77	2.48	0.13	0.88	1.10	1.20	1.37	2.40	0.19	0.28	85.5	85.7
Chemistry IB	Overfit	Lec_1_1	8	2012-2015	*Exam: Q1*													
				2014	Lec_1_1	423	362	-1.52	0.15	1.08	0.94	-0.70	0.75	-1.50	0.38	0.32	85.9	85.2
				2015	Lec_1_1	429	364	-1.55	0.14	1.06	0.96	-0.40	0.79	-1.30	0.34	0.29	84.9	84.6
		Lec_1_14	4	2012-2015	*Mirrored Molecule in Exam: Q14*													
				2012	Lec_1_14	384	291	-0.87	0.13	1.22	0.89	-1.90	0.74	-2.30	0.44	0.34	77.1	76.3
				2013	Lec_1_14	378	299	-0.98	0.14	1.24	0.84	-2.30	0.69	-2.60	0.48	0.34	82.8	79.1
		Lec_2_3	8	2012-2015	*Exam: Q17*													
				2012	Exam_2_17	421	252	-0.04	0.11	1.67	0.79	-5.70	0.73	-4.50	0.58	0.38	79.8	67.6
				2013	Lec_2_3	348	252	-1.05	0.13	1.23	0.89	-1.90	0.76	-2.30	0.49	0.38	78.0	75.1
				2013	Exam_2_17	434	278	-0.10	0.11	1.46	0.83	-4.20	0.74	-3.80	0.55	0.40	75.6	69.8
				2014	Lec_2_3	395	278	-0.92	0.12	1.44	0.79	-4.30	0.69	-3.40	0.55	0.38	79.4	73.1
				2014	Exam_2_17	456	301	-0.23	0.11	1.51	0.81	-4.80	0.72	-4.20	0.54	0.37	77.1	70.0
		Lec_2_7	2	2015	*Exam: Q21*													
				2015	Lec_2_7	393	353	-2.22	0.17	1.04	0.97	-0.20	0.80	-0.80	0.29	0.25	89.9	89.7
				2015	Exam_2_21	472	438	-2.31	0.18	1.04	0.98	-0.10	0.69	-1.20	0.24	0.20	92.7	92.7
		Exam_2_14	4	2012-2015	*Mirrored Molecule in Lecture 1: Q14*													
				2012	Exam_2_14	421	343	-1.28	0.13	1.12	0.92	-1.10	0.77	-1.70	0.39	0.29	83.6	81.7
				2013	Exam_2_14	434	347	-1.04	0.13	1.19	0.87	-1.90	0.74	-2.10	0.45	0.33	81.4	80.4
	Underfit	Lec_1_9	8	2012-2015	*Exam: Q9*													
				2012	Lec_1_9	384	191	0.51	0.11	0.37	1.19	4.30	1.27	4.30	0.27	0.43	58.8	67.5
				2013	Exam_2_9	434	258	0.14	0.11	0.46	1.17	4.10	1.29	3.90	0.26	0.42	63.3	68.5
				2015	Lec_1_9	429	228	0.34	0.11	0.48	1.16	3.80	1.25	4.00	0.31	0.44	61.2	67.8
				2015	Exam_2_9	472	262	0.34	0.10	0.53	1.14	3.50	1.26	3.80	0.31	0.43	64.6	68.1
		Lec_1_11	8	2012-2015	*Exam: Q11*													
				2012	Lec_1_11	384	230	0.00	0.11	0.57	1.13	3.00	1.21	2.90	0.29	0.41	60.4	67.6
				2014	Exam_2_11	456	266	0.16	0.10	0.22	1.22	5.70	1.32	4.70	0.19	0.39	57.6	67.0
		Lec_1_12	8	2012-2015	*Exam: Q12*													
				2013	Lec_1_12	378	173	0.83	0.12	0.59	1.13	2.80	1.21	3.30	0.34	0.45	64.8	68.3
				2014	Lec_1_12	423	193	0.91	0.11	0.24	1.28	5.60	1.45	6.10	0.29	0.50	60.0	70.3
				2015	Exam_2_12	472	273	0.22	0.10	0.43	1.18	4.50	1.27	3.80	0.27	0.42	59.2	68.2
		Lec_2_2	8	2012-2015	*Exam: Q23*													
				2013	Lec_2_2	348	107	1.13	0.13	0.82	1.09	1.40	1.20	1.90	0.35	0.43	70.8	74.4
				2015	Lec_2_2	393	98	1.70	0.13	0.80	1.14	1.90	1.25	2.10	0.34	0.45	77.1	79.5
		Lec_2_13	8	2012-2015	*Exam: Q29*													
				2012	Lec_2_13	364	219	-0.57	0.12	0.84	1.03	0.70	1.22	2.60	0.38	0.42	68.8	68.6
				2012	Exam_2_29	421	276	-0.33	0.11	0.70	1.09	2.10	1.29	3.40	0.25	0.36	67.2	69.9

			8	2012-2015														
		Exam_1_4		2012	Exam_1_4	434	84	1.59	0.13	0.77	1.20	2.40	1.31	2.10	0.24	0.39	80.0	82.1
				2013	Exam_1_4	450	110	1.55	0.12	0.74	1.15	2.30	1.49	3.50	0.29	0.42	75.7	78.8
				2015	Exam_1_4	487	126	1.31	0.12	0.72	1.18	3.00	1.29	2.70	0.28	0.42	72.4	77.6
			3	2012-2014														
		Exam_2_21		2013	Exam_2_21	434	243	0.31	0.11	0.48	1.17	4.00	1.23	3.50	0.28	0.43	61.2	68.3
				2014	Exam_2_21	456	284	-0.04	0.11	0.61	1.12	3.00	1.23	3.20	0.27	0.38	62.4	68.2
Foundations of Chemistry IA			4	2012-2015														
	Overfit	Lec_1_5		2013	Lec_1_5	309	276	-1.05	0.20	1.06	0.94	-0.40	0.79	-0.70	0.35	0.31	89.1	89.0
				2014	Lec_1_5	252	230	-1.39	0.24	1.12	0.84	-0.90	0.57	-1.30	0.43	0.33	91.9	91.1
			4	2012-2015														
		Lec_1_6		2012	Lec_1_6	259	234	-1.09	0.22	1.10	0.90	-0.60	0.62	-1.40	0.38	0.29	90.8	89.7
				2013	Lec_1_6	309	283	-1.34	0.22	1.05	0.96	-0.20	0.69	-1.00	0.32	0.28	91.5	91.3
				2014	Lec_1_6	252	225	-1.12	0.22	1.05	0.97	-0.10	0.74	-0.80	0.39	0.35	89.0	89.3
			4	2012-2015														
		Lec_1_9		2014	Lec_1_9	252	178	0.44	0.16	1.19	0.92	-1.20	0.79	-1.70	0.54	0.48	73.7	74.5
				2015	Lec_1_9	294	215	0.24	0.15	1.20	0.89	-1.50	0.79	-1.70	0.53	0.47	78.6	75.8
			4	2012-2015														
		Lec_1_10		2012	Lec_1_10	259	210	-0.19	0.17	1.14	0.92	-0.90	0.75	-1.60	0.45	0.37	81.5	80.5
				2015	Lec_1_10	294	257	-0.93	0.19	1.11	0.90	-0.80	0.68	-1.30	0.43	0.36	88.2	87.1
			4	2012-2015														
		Lec_2_1		2014	Lec_2_1	223	212	-2.88	0.32	1.10	0.87	-0.40	0.41	-1.40	0.33	0.21	95.1	95.1
				2015	Lec_2_1	236	221	-2.73	0.29	1.13	0.81	-0.90	0.44	-1.40	0.42	0.30	94.4	93.8
			4	2012-2015														
		Lec_2_2		2013	Lec_2_2	255	191	-0.98	0.16	1.16	0.91	-1.10	0.76	-1.40	0.50	0.43	79.4	78.2
				2014	Lec_2_2	223	176	-1.03	0.18	1.28	0.80	-2.20	0.61	-2.10	0.53	0.38	83.9	80.5
				2015	Lec_2_2	236	191	-1.23	0.19	1.15	0.86	-1.30	0.76	-1.10	0.50	0.42	85.9	83.1
			4	2012-2015														
		Lec_2_4		2012	Lec_2_4	267	228	-1.55	0.19	1.15	0.88	-1.10	0.61	-1.70	0.45	0.35	84.9	85.7
				2013	Lec_2_4	255	194	-1.06	0.16	1.19	0.88	-1.50	0.75	-1.40	0.51	0.42	81.8	79.0
				2015	Lec_2_4	236	191	-1.23	0.19	1.20	0.84	-1.50	0.64	-1.70	0.53	0.42	84.2	83.1
			4	2012-2015														
		Lec_2_6		2014	Lec_2_6	223	181	-1.20	0.19	1.06	0.97	-0.20	0.80	-0.90	0.40	0.36	83.4	82.2
				2015	Lec_2_6	236	173	-0.67	0.17	1.18	0.90	-1.20	0.74	-1.60	0.53	0.45	79.9	77.9
			4	2012-2015														
		Lec_2_14		2013	Lec_2_14	255	158	-0.20	0.15	1.36	0.85	-2.60	0.73	-2.60	0.58	0.47	77.9	72.1
				2015	Lec_2_14	236	196	-1.41	0.19	1.16	0.86	-1.20	0.67	-1.40	0.50	0.41	88.0	84.6
			4	2012-2015														
		Exam_1_7		2012	Exam_1_7	306	238	-1.63	0.15	1.16	0.89	-1.30	0.73	-1.80	0.49	0.40	83.9	79.8
				2013	Exam_1_7	365	275	-1.59	0.13	1.13	0.93	-1.00	0.80	-1.50	0.46	0.40	79.0	77.4
				2015	Exam_1_7	367	289	-1.76	0.14	1.20	0.86	-2.00	0.72	-1.90	0.50	0.39	81.9	80.3
			4	2012-2015														
		Exam_1_9		2012	Exam_1_9	306	198	-0.82	0.13	1.33	0.87	-2.50	0.76	-2.60	0.55	0.44	75.6	72.0
				2013	Exam_1_9	365	232	-0.89	0.12	1.42	0.83	-3.60	0.73	-3.20	0.57	0.44	79.0	71.5
				2014	Exam_1_9	327	209	-1.06	0.13	1.44	0.82	-3.60	0.69	-3.00	0.58	0.44	77.7	71.6
				2015	Exam_1_9	367	238	-0.89	0.12	1.39	0.83	-3.50	0.74	-3.00	0.57	0.45	79.1	72.3
			4	2012-2015														
		Exam_2_7		2012	Exam_2_7	258	242	-2.37	0.27	1.05	0.95	-0.20	0.64	-1.00	0.30	0.22	93.8	93.8
				2013	Exam_2_7	346	320	-2.59	0.27	1.04	0.96	-0.10	0.62	-1.10	0.57	0.54	95.5	95.5
			4	2012-2015														
		Exam_2_9		2012	Exam_2_9	258	215	-1.14	0.18	1.17	0.87	-1.20	0.62	-2.20	0.47	0.33	83.3	83.8
				2013	Exam_2_9	346	271	-0.86	0.15	1.18	0.88	-1.50	0.68	-2.00	0.53	0.44	82.4	81.2
				2014	Exam_2_9	300	204	-0.17	0.14	1.25	0.89	-2.00	0.80	-1.70	0.48	0.39	75.9	72.5
				2015	Exam_2_9	331	221	-0.15	0.13	1.41	0.83	-3.40	0.72	-3.00	0.55	0.41	77.3	72.2
			4	2012-2015														
		Exam_2_10		2012	Exam_2_10	258	238	-2.11	0.24	1.06	0.92	-0.40	0.78	-0.60	0.32	0.24	92.6	92.2
				2013	Exam_2_10	346	307	-1.89	0.20	1.07	0.94	-0.40	0.63	-1.40	0.54	0.49	91.6	91.6
			4	2012-2015														
		Exam_2_30		2013	Exam_2_30	347	185	0.59	0.12	1.46	0.86	-3.40	0.80	-2.80	0.50	0.42	73.2	67.3
				2014	Exam_2_30	299	155	0.68	0.13	1.48	0.84	-3.50	0.79	-2.70	0.56	0.44	74.2	68.4
				2015	Exam_2_30	328	200	0.18	0.13	1.46	0.84	-3.50	0.74	-3.30	0.56	0.42	73.4	69.4
			4	2012-2015														
	Underfit	Lec_1_1		2013	Lec_1_1	309	283	-1.34	0.22	0.94	1.07	0.50	1.31	1.00	0.22	0.28	91.5	91.3
				2014	Lec_1_1	252	231	-1.45	0.25	0.87	1.12	0.70	1.70	1.70	0.22	0.33	90.7	91.5
			4	2012-2015														
		Lec_1_4		2012	Lec_1_4	259	93	2.46	0.16	0.67	1.18	2.30	1.25	2.00	0.41	0.53	71.0	74.4
				2014	Lec_1_4	252	74	2.90	0.17	0.77	1.12	1.40	1.70	3.20	0.43	0.52	77.1	78.6
				2015	Lec_1_4	294	94	2.73	0.16	0.73	1.14	1.80	1.53	3.00	0.46	0.56	74.9	78.2
			4	2012-2015														
		Lec_2_8		2012	Lec_2_8	267	50	2.59	0.18	0.81	1.08	0.80	1.69	2.80	0.31	0.43	84.5	83.0
				2013	Lec_2_8	255	37	2.74	0.20	0.91	1.02	0.20	1.42	1.50	0.34	0.40	87.4	86.9
				2014	Lec_2_8	223	48	2.27	0.18	0.78	1.10	1.00	1.75	3.30	0.30	0.43	82.1	81.5
				2015	Lec_2_8	236	54	2.25	0.18	0.74	1.11	1.20	1.86	3.20	0.31	0.44	81.2	80.8
			4	2012-2015														
		Lec_2_9		2013	Lec_2_9	255	206	-1.40	0.18	0.81	1.15	1.50	1.32	1.40	0.29	0.39	80.6	82.2
				2015	Lec_2_9	236	193	-1.30	0.19	0.79	1.16	1.40	1.53	2.00	0.29	0.41	81.6	83.7
			4	2012-2015														
		Lec_2_13		2012	Lec_2_13	267	44	2.79	0.18	0.70	1.26	2.20	1.56	2.10	0.21	0.41	83.8	85.1
				2013	Lec_2_13	255	41	2.59	0.19	0.46	1.41	3.20	2.78	4.90	0.06	0.41	84.2	85.4
				2014	Lec_2_13	223	30	2.97	0.22	0.66	1.29	2.00	2.31	3.50	0.11	0.38	85.7	87.5
				2015	Lec_2_13	236	40	2.73	0.20	0.69	1.25	2.00	2.11	3.20	0.22	0.41	82.9	85.0
			4	2012-2015														
		Lec_2_15		2014	Lec_2_15	223	61	1.87	0.17	0.92	1.01	0.20	1.33	2.00	0.41	0.45	76.7	77.6
				2015	Lec_2_15	236	75	1.67	0.16	0.89	1.02	0.40	1.41	2.30	0.43	0.47	76.5	75.5
			4	2012-2015														
		Exam_1_6		2012	Exam_1_6	306	73	1.43	0.15	0.86	1.06	0.80	1.50	3.00	0.36	0.44	78.9	80.0
				2013	Exam_1_6	365	101	1.09	0.13	0.83	1.08	1.20	1.28	2.30	0.40	0.47	79.6	77.7
				2014	Exam_1_6	327	72	1.35	0.15	0.80	1.14	1.50	1.52	2.90	0.36	0.47	80.8	82.2
			4	2012-2015														
		Exam_1_10		2013	Exam_1_10	365	50	2.24	0.17	0.88	1.11	1.00	1.40	1.70	0.33	0.43	86.3	88.1
				2014	Exam_1_10	327	29	2.74	0.22	0.82	1.12	0.80	2.61	3.50	0.23	0.39	91.6	91.8
				2015	Exam_1_10	367	50	2.31	0.17	0.84	1.17	1.40	1.39	1.60	0.30	0.42	85.5	88.4
			4	2012-2015														
		Exam_2_1		2012	Exam_2_1	258	244	-2.52	0.28	1.01	0.97	-0.10	1.25	0.70	0.23	0.21	94.6	94.5
				2013	Exam_2_1	346	313	-2.17	0.23	0.94	1.03	0.20	2.00	2.50	0.46	0.51	93.4	93.4
				2014	Exam_2_1	300	274	-2.01	0.22	0.96	1.00	0.00	1.30	1.00	0.21	0.25	91.6	91.5
			4	2012-2015														
		Exam_2_4		2013	Exam_2_4	346	271	-0.86	0.15	0.77	1.16	1.90	1.36	2.00	0.34	0.44	78.2	81.2
				2014	Exam_2_4	300	209	-0.26	0.14	0.57	1.19	3.00	1.36	2.50	0.23	0.39	67.6	73.4
				2015	Exam_2_4	331	250	-0.66	0.14	0.87	1.06	0.90	1.25	1.70	0.31	0.37	77.6	77.6
			4	2012-2015														
		Exam_2_12		2013	Exam_2_12	346	176	0.71	0.12	0.42	1.16	3.60	1.31	4.00	0.31	0.42	59.1	67.3
				2015	Exam_2_12	331	193	0.30	0.12	0.64	1.11	2.30	1.21	2.40	0.32	0.43	65.2	68.9
			4	2012-2015														
		Exam_2_18		2012	Exam_2_18	258	81	1.72	0.15	0.63	1.16	2.40	1.33	2.80	0.24	0.40	70.4	73.4
				2013	Exam_2_18	347	78	2.31	0.14	0.76	1.15	1.80	1.38	2.90	0.25	0.36	78.3	80.0
				2014	Exam_2_18	299	65	2.39	0.16	0.64	1.29	3.00	1.59	3.20	0.21	0.44	77.2	81.8
				2015	Exam_2_18	328	69	2.37	0.15	0.78	1.15	1.70	1.46	2.80	0.26	0.41	80.1	81.2
			4	2012-2015														
		Exam_2_29		2012	Exam_2_29	258	50	2.49	0.17	0.87	1.07	0.70	1.29	1.60	0.27	0.36	80.9	81.6
				2013	Exam_2_29	347	47	3.06	0.17	0.82	1.13	1.10	1.77	3.40	0.20	0.32	87.2	87.2
				2014	Exam_2_29	299	64	2.41	0.16	0.54	1.30	3.00	3.06	8.50	0.14	0.43	76.8	82.0
				2015	Exam_2_29	328	58	2.64	0.16	0.65	1.26	2.60	1.96	4.40	0.13	0.40	82.6	83.9
Foundations of Chemistry IB			4	2012-2015														
	Overfit	Lec_1_4		2012	Lec_1_4	238	180	-0.39	0.17	1.21	0.88	-1.40	0.68	-1.90	0.53	0.45	81.3	78.2
				2014	Lec_1_4	216	161	-0.57	0.18	1.27	0.84	-1.90	0.65	-2.20	0.57	0.46	81.0	78.3
			4	2012-2015														
		Lec_1_6		2012	Lec_1_6	238	206	-1.30	0.21	1.09	0.92	-0.60	0.72	-0.90	0.42	0.36	87.1	86.5

			2013	Lec_1_6	249	223	-1.86	0.22	1.13	0.84	-1.10	0.69	-0.90	0.41	0.31	89.3	89.3
			2014	Lec_1_6	216	190	-1.72	0.23	1.15	0.83	-1.10	0.57	-1.40	0.47	0.36	91.0	88.4
			2015	Lec_1_6	231	193	-1.41	0.19	1.17	0.87	-1.10	0.60	-1.80	0.47	0.37	83.3	84.2
		4	2012-2015														
	Lec_1_7		2014	Lec_1_7	216	112	0.77	0.16	1.39	0.83	-2.60	0.75	-2.40	0.62	0.52	78.6	71.8
			2015	Lec_1_7	231	88	1.29	0.16	1.29	0.87	-1.90	0.75	-2.20	0.59	0.49	73.2	73.1
		4	2012-2015	*Lecture 2 (2013 - 2015): Q4*													
	Lec_1_13		2012	Lec_1_13	238	202	-1.14	0.20	1.14	0.88	-1.00	0.67	-1.20	0.46	0.38	86.2	85.1
			2013	Lec_2_4	218	174	-1.24	0.19	1.37	0.73	-2.90	0.47	-3.00	0.60	0.39	84.8	81.0
			2014	Lec_2_4	184	140	-0.65	0.20	1.29	0.81	-2.00	0.70	-1.50	0.54	0.43	81.8	78.8
			2015	Lec_2_4	198	148	-0.61	0.18	1.25	0.87	-1.60	0.68	-1.80	0.52	0.42	79.4	77.3
	Lec_2_8	1	2012														
			2012	Lec_2_8	189	110	-0.38	0.17	1.34	0.87	-2.00	0.75	-2.10	0.57	0.48	75.7	71.2
		4	2012-2015	*Lecture 2 (2014-2015): Q13*													
	Lec_2_12		2012	Lec_2_12	189	79	0.50	0.17	1.27	0.89	-1.60	0.80	-1.70	0.57	0.49	75.1	72.3
			2014	Lec_2_13	184	77	1.36	0.18	1.31	0.85	-1.90	0.75	-2.00	0.63	0.54	80.1	73.6
	Exam_1_8	1	2012														
			2012	Exam_1_8	266	225	-1.22	0.18	1.16	0.89	-1.00	0.59	-2.10	0.46	0.35	82.5	84.7
		4	2012-2015														
	Exam_2_4		2012	Exam_2_4	249	224	-1.91	0.22	1.08	0.90	-0.60	0.78	-0.60	0.34	0.28	90.1	89.8
			2013	Exam_2_4	286	255	-1.59	0.20	1.04	0.97	-0.20	0.80	-0.60	0.31	0.29	89.0	89.0
			2014	Exam_2_4	259	219	-1.51	0.18	1.14	0.89	-1.00	0.67	-1.40	0.40	0.31	86.0	84.7
			2015	Exam_2_4	277	237	-1.51	0.18	1.15	0.86	-1.30	0.66	-1.50	0.44	0.32	86.6	85.9
	Exam_2_8	4	2012-2015														
			2013	Exam_2_8	286	236	-0.95	0.17	1.14	0.90	-1.10	0.73	-1.20	0.43	0.35	83.3	82.6
			2014	Exam_2_8	259	204	-1.06	0.16	1.11	0.94	-0.70	0.79	-1.10	0.41	0.35	81.4	79.6
	Exam_2_11	4	2012-2015	*Exam (2013-2015): Q17*													
			2013	Exam_2_17	287	214	-0.37	0.15	1.20	0.88	-1.70	0.79	-1.30	0.49	0.41	80.5	76.8
			2014	Exam_2_17	259	175	-0.38	0.15	1.41	0.80	-3.50	0.80	-1.60	0.54	0.40	79.8	72.2
			2015	Exam_2_17	277	189	-0.31	0.14	1.47	0.77	-4.00	0.66	-3.20	0.60	0.41	80.9	73.5
	Exam_2_13	4	2012-2015	*Exam (2013-2015): Q19 Stem Change*													
			2013	Exam_2_19	287	164	0.64	0.14	1.36	0.85	-2.70	0.78	-2.50	0.58	0.48	77.3	71.4
			2015	Exam_2_19	277	153	0.38	0.14	1.47	0.83	-3.40	0.76	-3.00	0.58	0.44	76.2	69.4
	Exam_2_15	4	2012-2015	*Exam (2013-2015): Q21*													
			2012	Exam_2_15	249	188	-0.66	0.16	1.23	0.86	-1.80	0.72	-1.70	0.49	0.39	82.2	77.5
			2013	Exam_2_21	287	211	-0.30	0.15	1.21	0.88	-1.70	0.76	-1.60	0.50	0.41	79.8	76.3
			2014	Exam_2_21	259	196	-0.86	0.16	1.17	0.91	-1.20	0.76	-1.40	0.44	0.36	76.7	77.2
			2015	Exam_2_21	277	198	-0.50	0.15	1.16	0.94	-0.90	0.78	-1.70	0.47	0.40	72.6	75.2
Underfit		4	2012-2015														
	Lec_1_1		2012	Lec_1_1	238	210	-1.49	0.22	0.90	1.11	0.80	1.26	0.80	0.28	0.35	87.1	88.1
			2014	Lec_1_1	216	183	-1.38	0.21	0.86	1.12	0.90	1.37	1.30	0.31	0.39	82.9	85.6
			2015	Lec_1_1	231	203	-1.83	0.22	1.06	0.88	-0.80	1.51	1.50	0.38	0.33	88.6	88.2
		4	2012-2015														
	Lec_1_2		2012	Lec_1_2	238	52	3.06	0.20	0.66	1.31	2.60	1.58	2.00	0.38	0.54	77.8	84.3
			2013	Lec_1_2	249	55	2.55	0.18	0.88	1.09	0.90	1.32	1.50	0.41	0.48	80.7	82.0
		4	2012-2015														
	Lec_1_3		2012	Lec_1_3	238	214	-1.69	0.23	0.81	1.14	0.90	2.13	2.40	0.19	0.33	88.9	89.7
			2014	Lec_1_3	216	189	-1.67	0.23	0.93	1.02	0.20	1.34	1.00	0.32	0.37	89.5	88.0
			2015	Lec_1_3	231	200	-1.70	0.21	0.94	1.01	0.20	1.29	1.00	0.30	0.34	87.3	87.0
		4	2012-2015														
	Lec_1_8		2013	Lec_1_8	249	199	-0.95	0.18	0.87	1.09	0.90	1.21	1.00	0.33	0.40	81.1	81.2
			2014	Lec_1_8	216	168	-0.80	0.19	0.84	1.09	1.00	1.26	1.20	0.37	0.44	79.5	80.4
	Lec_2_7	1	2012														
			2012	Lec_2_7	189	89	0.21	0.17	0.63	1.14	2.00	1.27	2.20	0.38	0.49	64.0	71.0
		2	2014-2015														
	Lec_2_11		2014	Lec_2_11	184	176	-2.94	0.38	0.86	1.17	0.60	2.28	1.70	0.10	0.23	95.5	95.4
			2015	Lec_2_11	198	189	-2.86	0.35	0.95	0.96	0.00	2.12	1.70	0.16	0.21	95.4	95.4
		3	2013-2015	*Lecture 2 (2014-2015): Q14*													
	Lec_2_13		2013	Lec_2_13	218	103	0.66	0.16	0.50	1.23	3.20	1.29	2.80	0.33	0.49	61.8	71.9
			2015	Lec_2_14	198	92	0.99	0.17	0.61	1.18	2.40	1.21	1.90	0.40	0.51	63.9	71.3
		4	2012-2015	*Exam 1 (2013-2015): Q5*													
	Exam_1_6		2013	Exam_1_5	305	95	0.47	0.14	0.80	1.10	1.50	1.20	1.80	0.38	0.46	70.4	74.7
			2014	Exam_1_5	276	73	1.35	0.16	0.84	1.08	1.00	1.39	2.50	0.39	0.47	76.3	78.7
		7	2012-2015	*Lecture 2 (2013-2015): Q8, Exam (2013-2015): Q23*													
	Exam_1_3		2012	Exam_1_3	266	96	1.66	0.15	0.61	1.18	2.70	1.28	2.40	0.35	0.49	71.2	73.4
			2014	Exam_2_23	259	62	2.03	0.17	0.72	1.22	2.30	1.33	2.00	0.30	0.46	74.8	80.5
			2015	Lec_2_8	198	62	1.87	0.18	0.64	1.20	2.10	1.48	2.70	0.36	0.51	71.1	76.8
		2	2014-2015														
	Exam_1_9		2014	Exam_1_9	276	221	-1.68	0.16	0.94	1.00	0.00	1.59	2.90	0.32	0.35	81.5	80.8
			2015	Exam_1_9	300	251	-1.87	0.17	0.94	1.02	0.20	1.23	1.10	0.30	0.35	84.1	84.1
		4	2012-2015														
	Exam_2_2		2013	Exam_2_2	286	267	-2.17	0.25	0.93	1.02	0.10	3.98	4.60	0.16	0.23	93.2	93.2
			2015	Exam_2_2	277	252	-2.11	0.22	0.85	1.14	0.90	1.78	2.10	0.11	0.27	91.0	91.0
		4	2012-2015														
	Exam_2_9		2012	Exam_2_9	249	68	2.08	0.17	0.79	1.11	1.30	1.46	2.70	0.41	0.50	77.7	79.0
			2013	Exam_2_9	286	65	2.66	0.16	0.80	1.14	1.50	1.34	1.90	0.38	0.49	79.7	81.3
			2014	Exam_2_9	259	64	1.97	0.16	0.74	1.21	2.30	1.25	1.60	0.32	0.46	74.0	80.0
			2015	Exam_2_9	277	62	2.20	0.16	0.90	1.05	0.60	1.24	1.50	0.35	0.41	80.5	80.4
		4	2012-2015	*Exam (2013-2015): Q20*													
	Exam_2_14		2012	Exam_2_14	249	182	-0.51	0.16	0.69	1.15	1.90	1.46	2.60	0.28	0.40	72.7	76.1
			2015	Exam_2_20	277	205	-0.65	0.15	0.64	1.19	2.40	1.69	3.90	0.22	0.39	72.9	76.9
		4	2012-2015	*Exam (2013-2015): Q15*													
	Exam_2_21		2012	Exam_2_21	277	198	-0.50	0.15	1.16	0.94	-0.90	0.78	-1.70	0.47	0.40	72.6	75.2
			2013	Exam_2_15	286	84	2.19	0.15	0.54	1.24	3.00	1.67	4.30	0.30	0.50	74.0	77.2
			2014	Exam_2_15	259	64	1.97	0.16	0.88	1.04	0.50	1.40	2.50	0.40	0.46	79.5	80.0
			2015	Exam_2_15	277	76	1.86	0.15	0.65	1.16	2.00	1.59	3.90	0.25	0.43	76.5	76.9

7.10 Comparison of Male and Female Raw Scores in Chemistry MCQ Assessments

Table 75: The Raw Score Comparison of Male and Female Student Results in Chemistry IA MCQ Assessments to Determine if a Difference Should be Expected in how each Gender Answers Individual Items (All Significance Favours Male Students)

Chemistry IA	2012		2013		2014		2015	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	466	0.006	445	0.223	470	0.266	502	0.02
Lecture Test 2	444	0.175	418	0.006	434	0.029	449	<<0.001
Exam	499	0.01	499	0.319	505	0.212	538	0.377
Redeemable Exam	485	0.977	486	0.546	496	0.021	523	0.237
Percentage	516	0.099	517	0.163	524	0.006	558	0.004

Table 76: The Raw Score Comparison of Male and Female Student Results in Chemistry IB MCQ Assessments to Determine if a Difference Should be Expected in how each Gender Answers Individual Items (All Significance Favours Male Students)

Chemistry IB	2012		2013		2014		2015	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	380	0.203	376	0.072	421	0.972	426	0.002
Lecture Test 2	361	0.595	346	0.01	392	0.06	389	0.005
Exam	425	0.847	446	0.879	479	0.07	478	0.509
Redeemable Exam	419	0.567	432	0.81	454	0.511	469	0.706
Percentage	443	0.077	451	0.547	488	0.049	490	0.551

Table 77: The Raw Score Comparison of Male and Female Student Results in Foundations of Chemistry IA MCQ Assessments to Determine if a Difference Should be Expected in how each Gender Answers Individual Items

Foundations of Chemistry IA	2012		2013		2014		2015	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	257	0.925	307	0.902	250	0.922	292	0.223
Lecture Test 2	265	0.132	253	0.723	221	0.473	234	0.693
Exam	301	0.315	360	0.611	323	0.211	363	0.757
Redeemable Exam	256	0.926	334	0.157	299	0.251	329	0.539
Percentage	321	0.193	377	0.112	340	0.726	385	0.149

Table 78: The Raw Score Comparison of Male and Female Student Results in Foundations of Chemistry IB MCQ Assessments to Determine if a Difference Should be Expected in how each Gender Answers Individual Items (2012 Favours Female Students, 2015 Favours Male Students)

Foundations of Chemistry IB	2012		2013		2014		2015	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	234	0.604	247	0.935	214	0.961	229	0.036
Lecture Test 2	187	0.042	216	0.864	182	0.947	196	0.304
Exam	264	0.125	296	0.746	274	0.694	297	0.075
Redeemable Exam	248	0.376	286	0.204	257	0.864	275	0.29
Percentage	280	0.722	315	0.91	283	0.922	306	0.453

7.11 Boxplot Comparison of Male and Female Raw Score in Chemistry MCQ Assessments

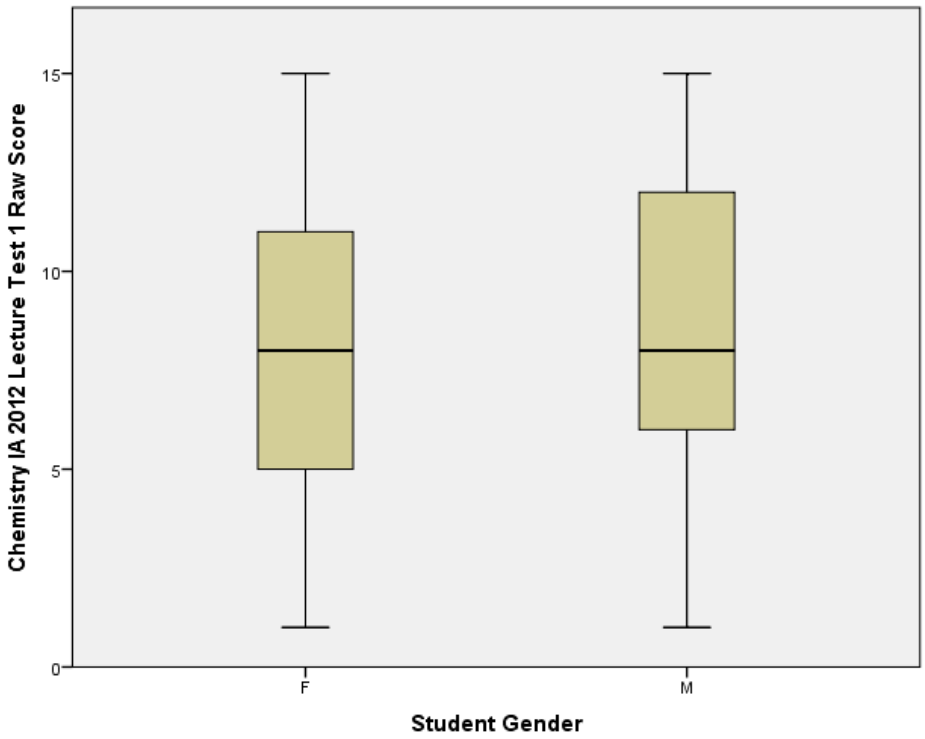

Figure 365: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Lecture Test 1 2012 to Observe Significant Differences

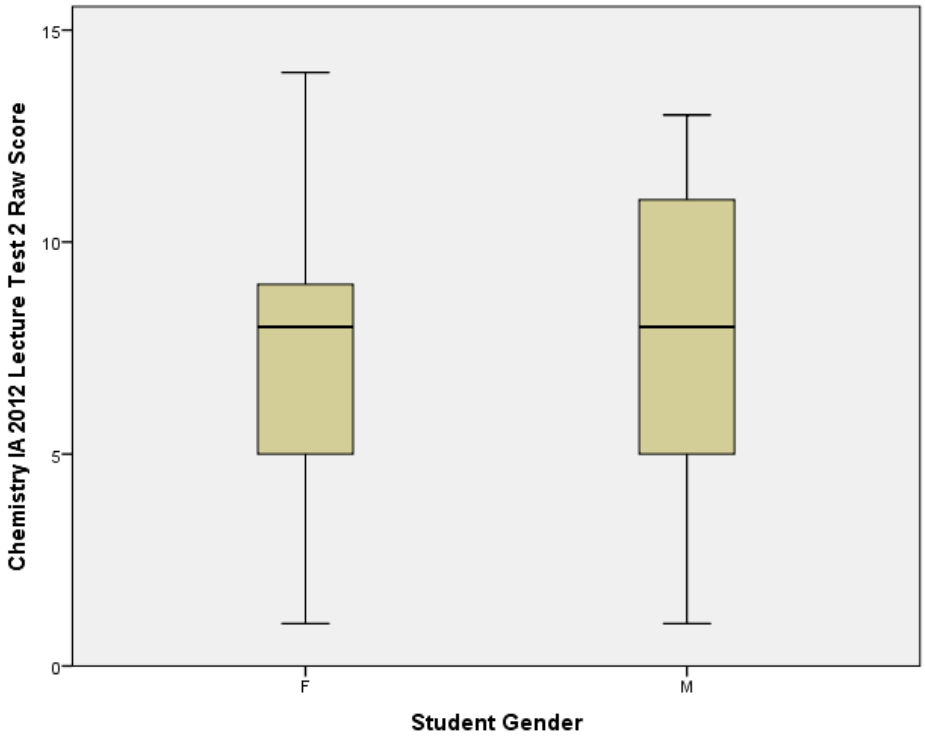

Figure 366: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Lecture Test 2 2012 to Observe Significant Differences

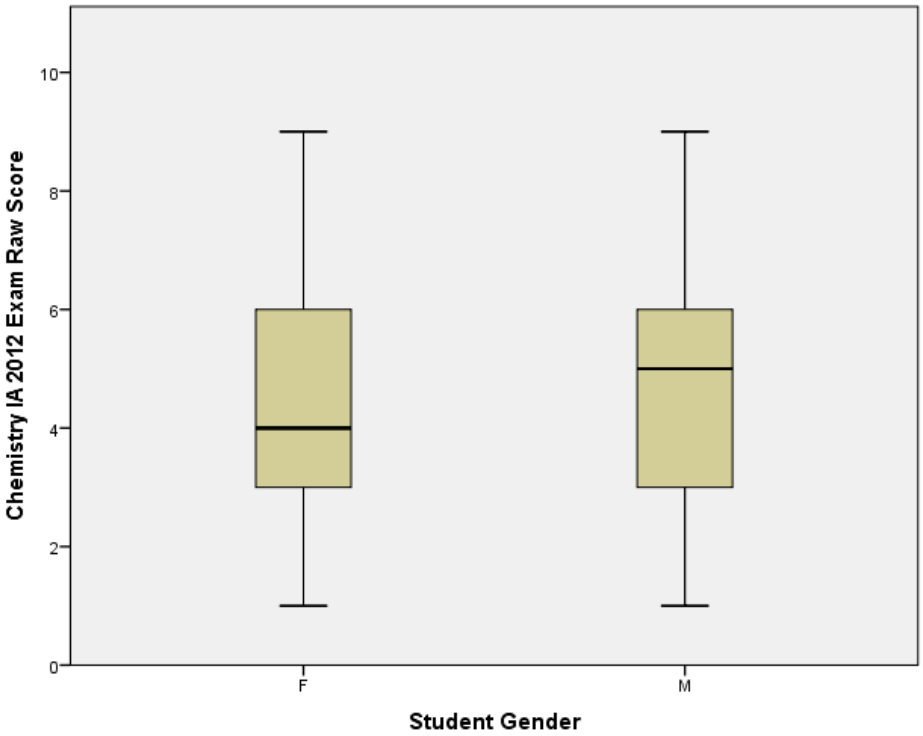

Figure 367: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Exam 2012 to Observe Significant Differences

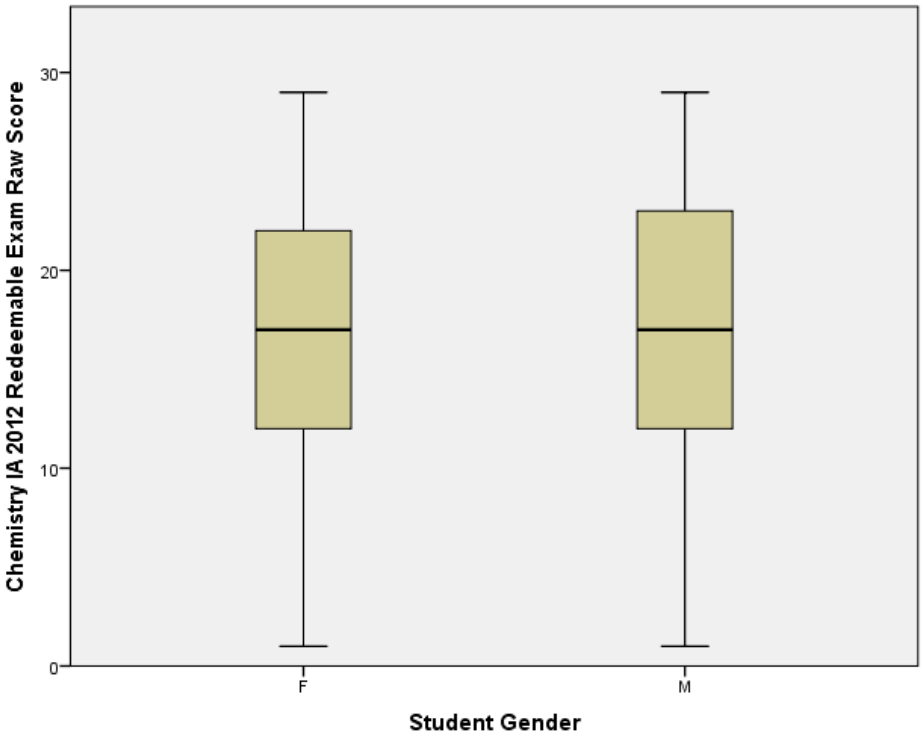

Figure 368: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Redeemable Exam 2012 to Observe Significant Differences

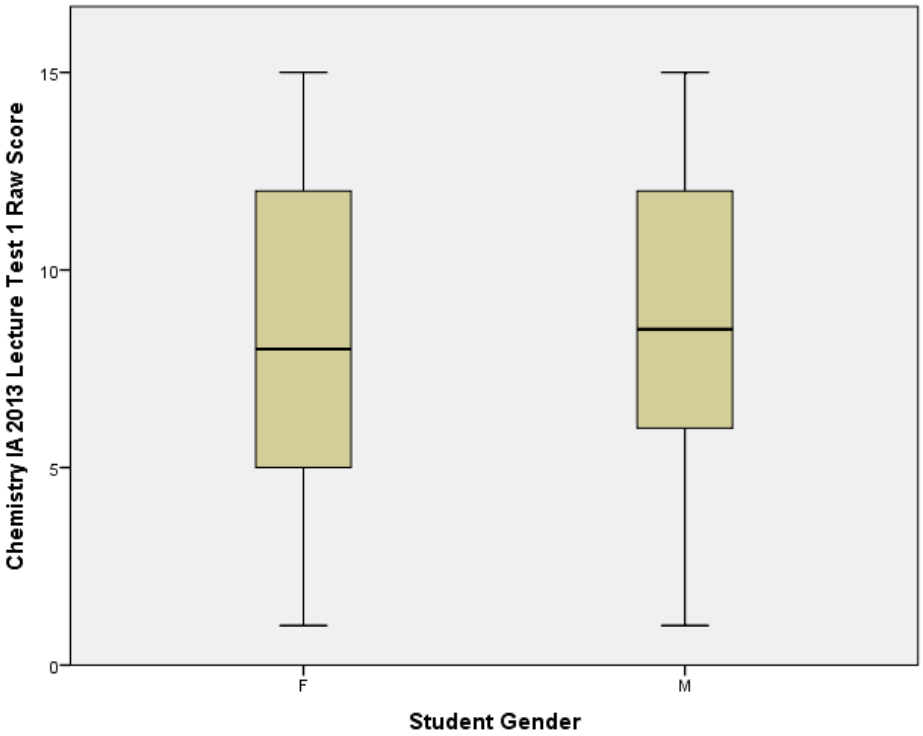

Figure 369: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Lecture Test 1 2013 to Observe Significant Differences

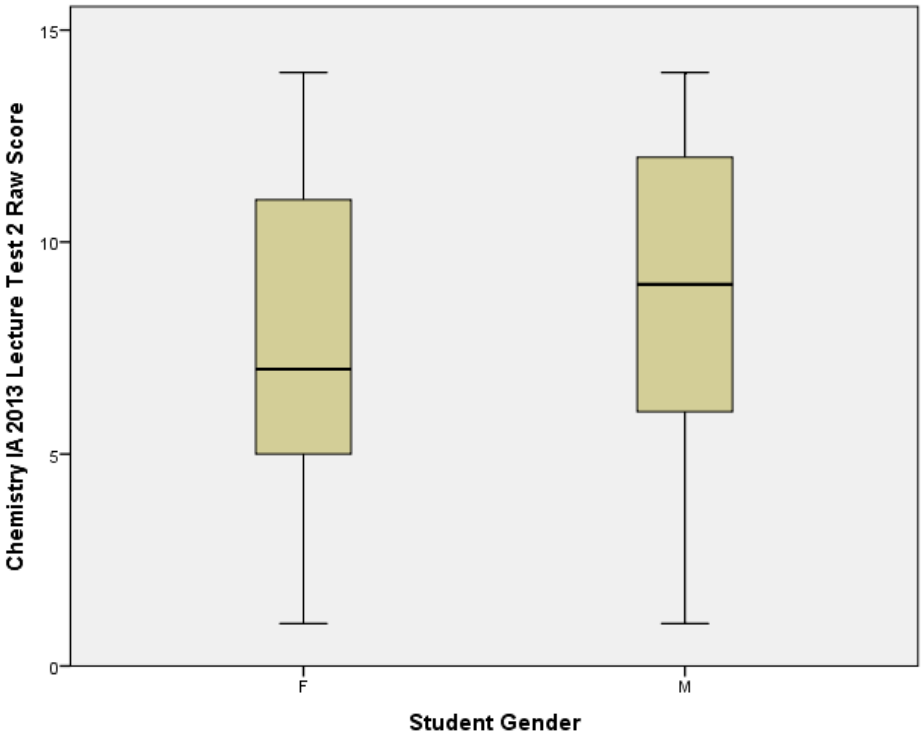

Figure 370: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Lecture Test 2 2013 to Observe Significant Differences

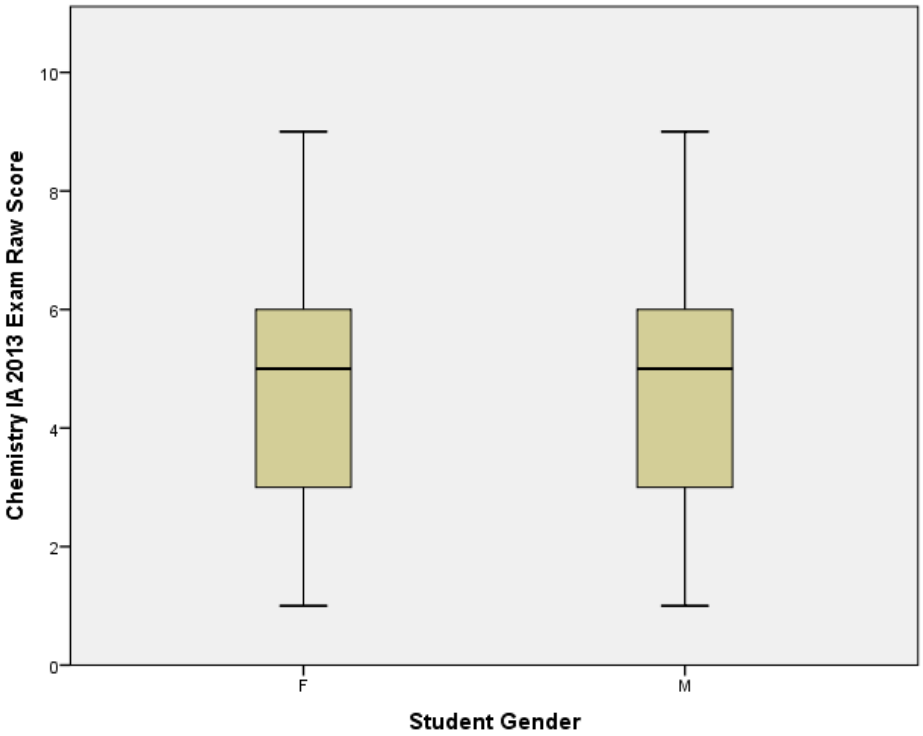

Figure 371: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Exam 2013 to Observe Significant Differences

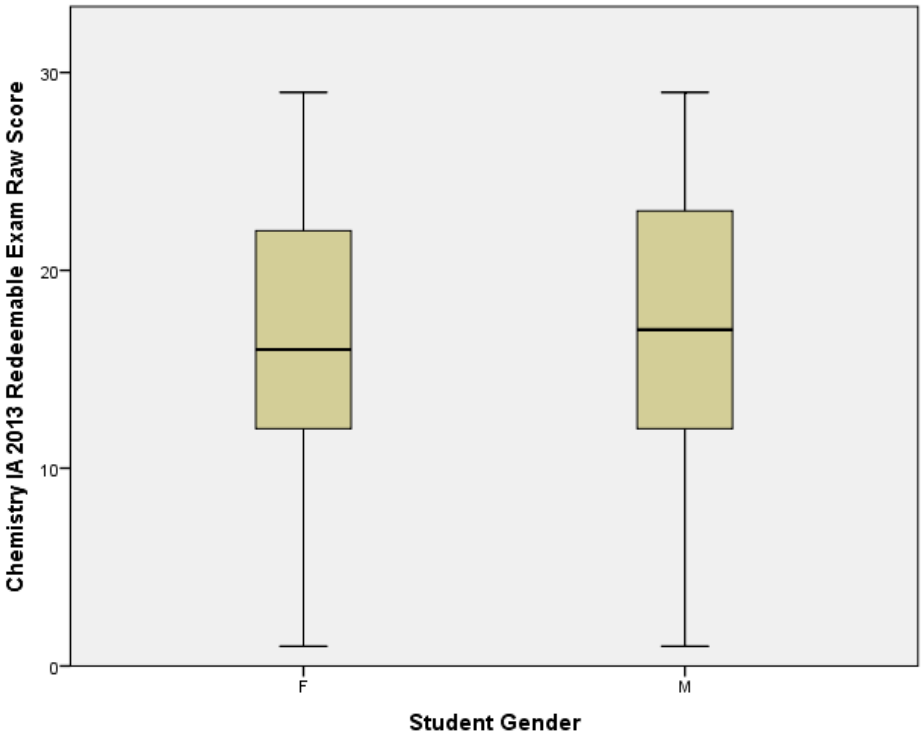

Figure 372: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Redeemable Exam 2013 to Observe Significant Differences

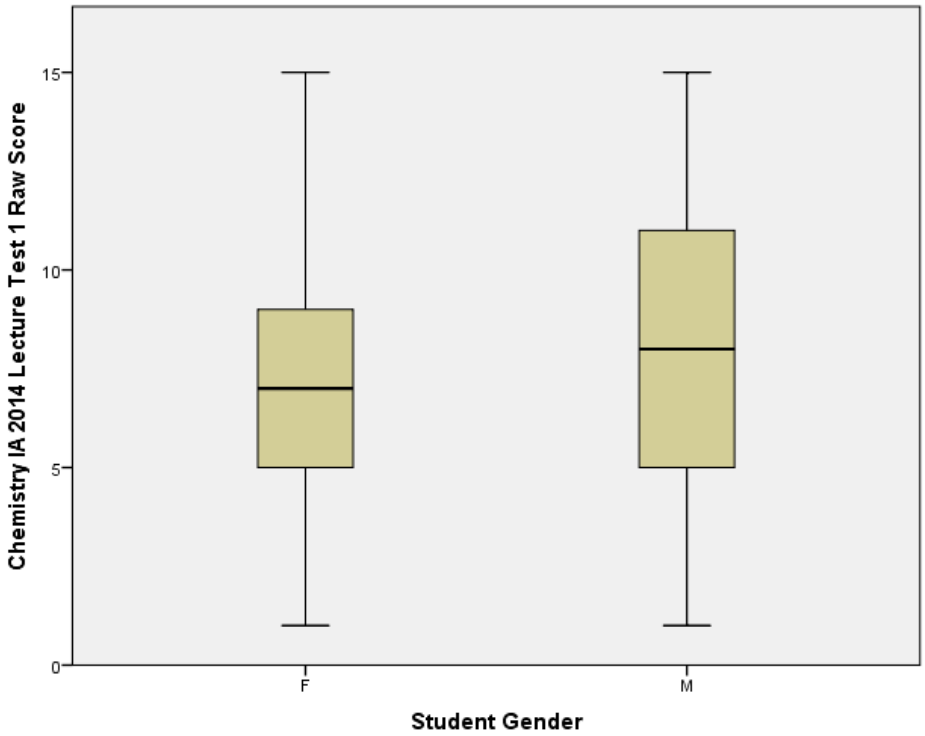

Figure 373: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Lecture Test 1 2014 to Observe Significant Differences

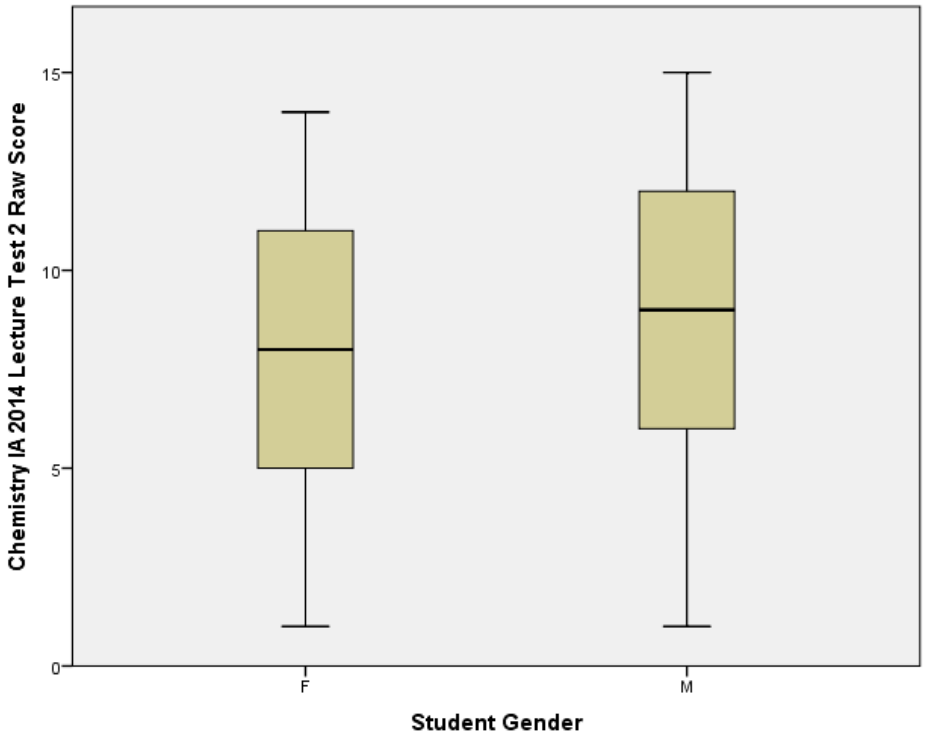

Figure 374: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Lecture Test 2 2014 to Observe Significant Differences

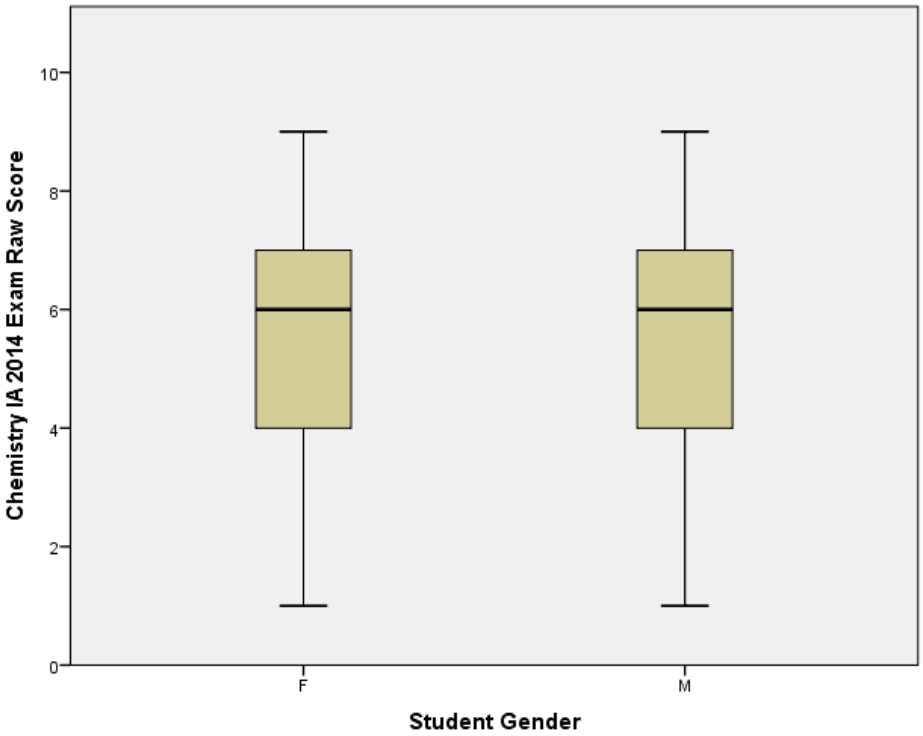

Figure 375: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Exam 2014 to Observe Significant Differences

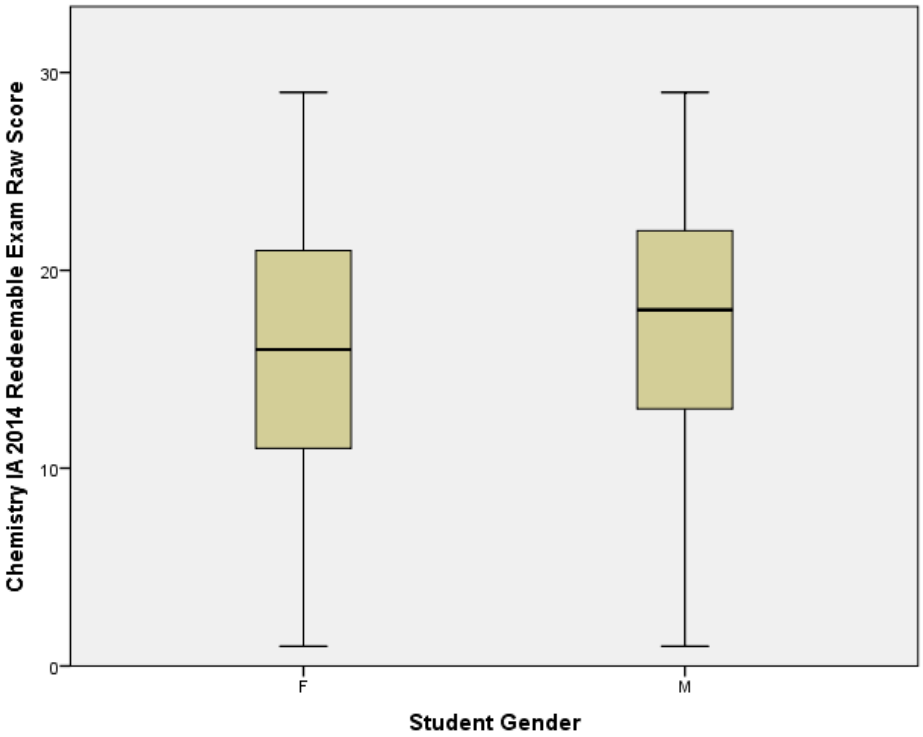

Figure 376: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Redeemable Exam 2014 to Observe Significant Differences

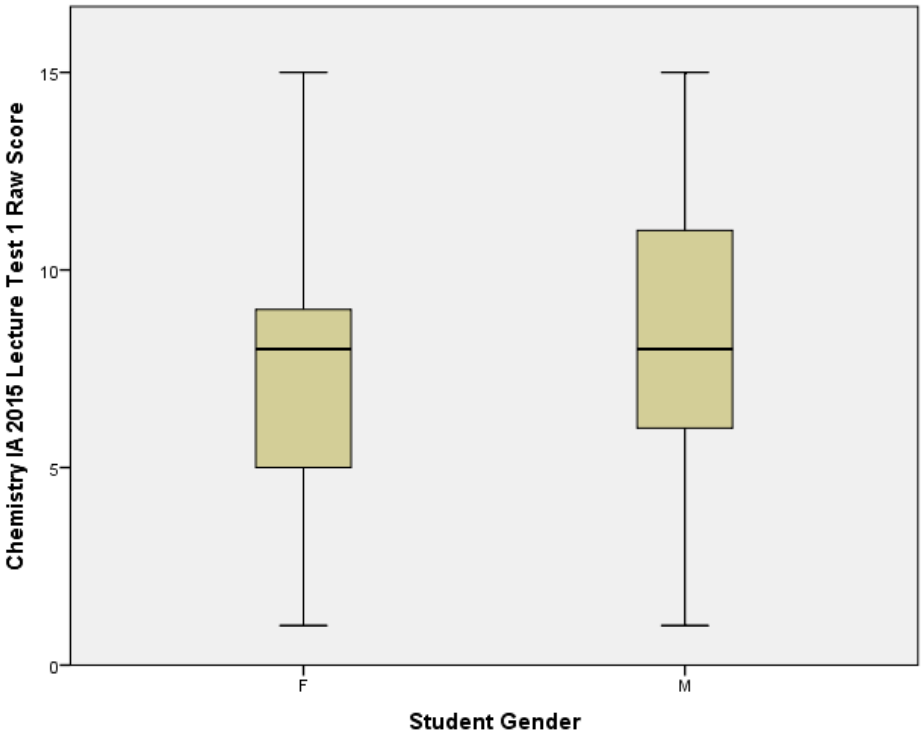

Figure 377: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Lecture Test 1 2015 to Observe Significant Differences

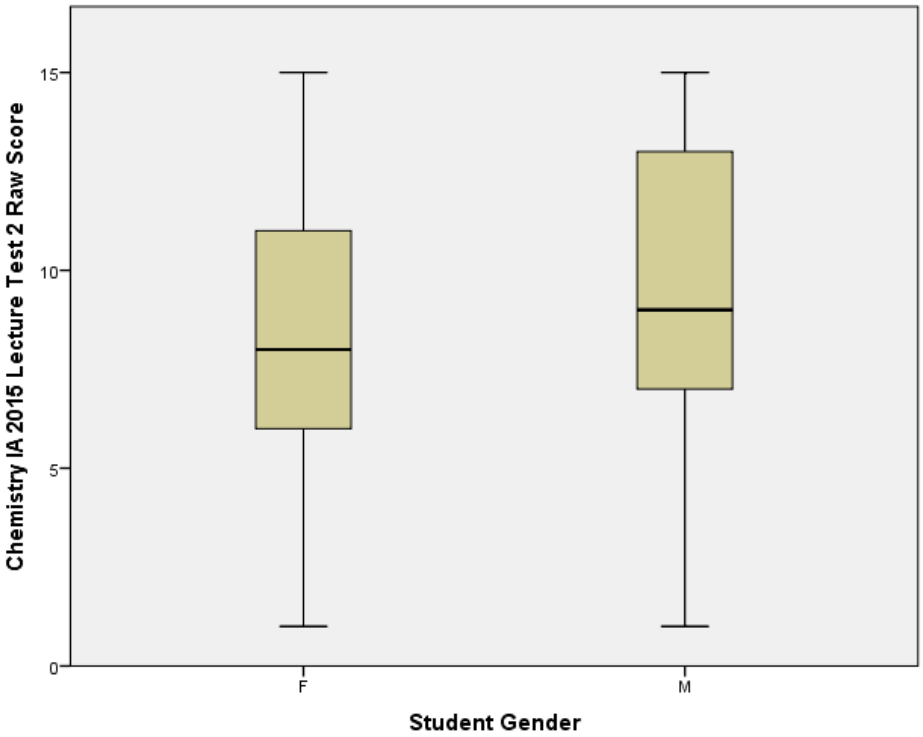

Figure 378: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Lecture Test 2 2015 to Observe Significant Differences

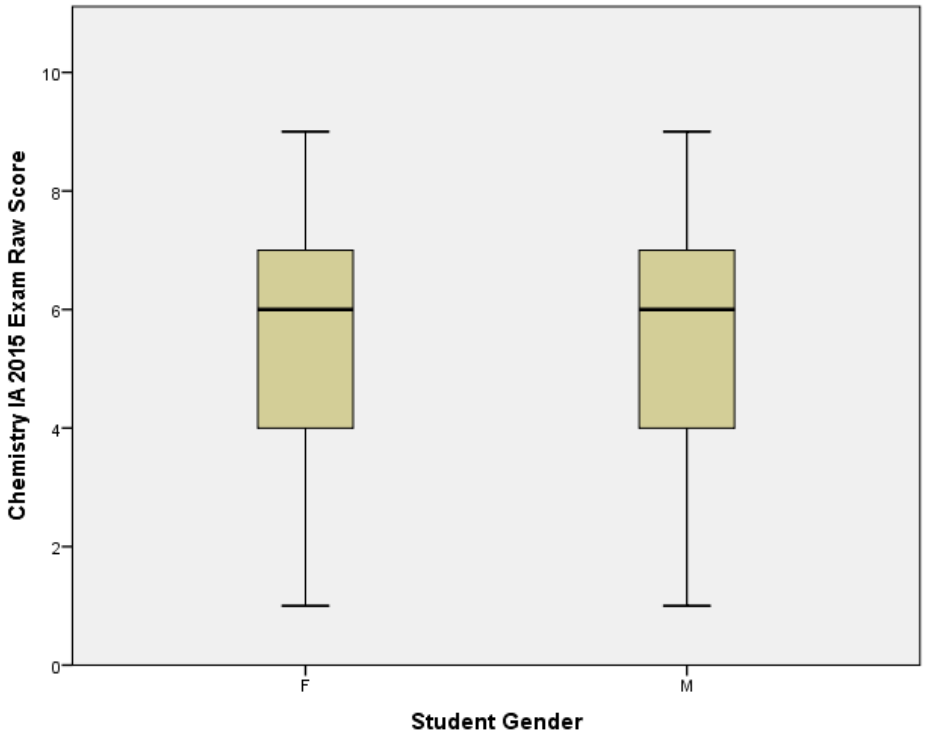

Figure 379: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Exam 2015 to Observe Significant Differences

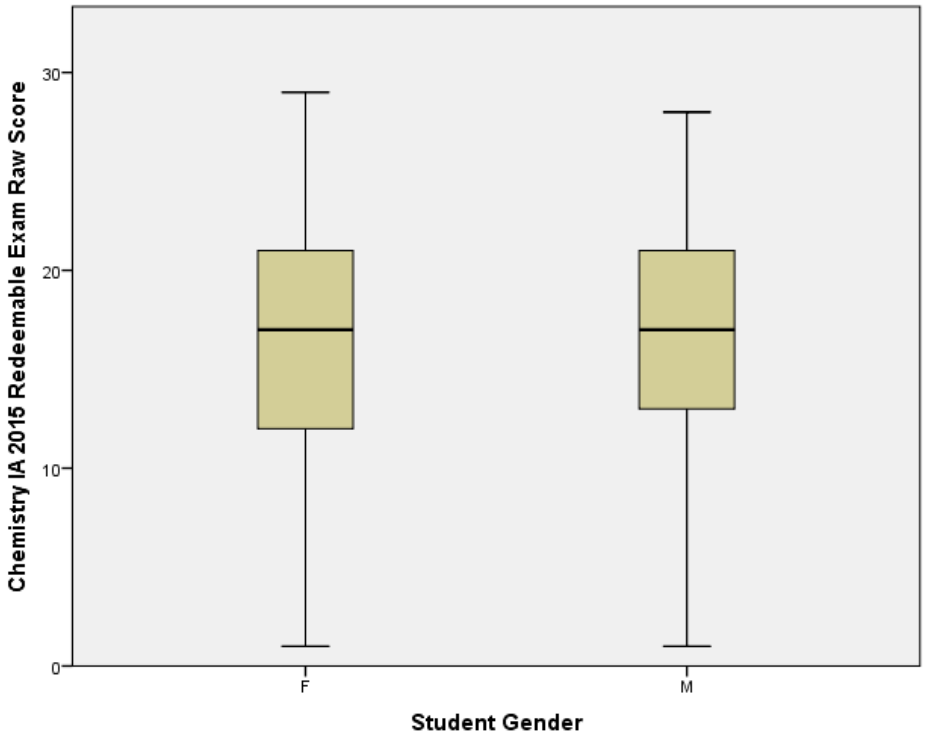

Figure 380: The Boxplot Comparison of Male and Female Raw Score in Chemistry IA Redeemable Exam 2015 to Observe Significant Differences

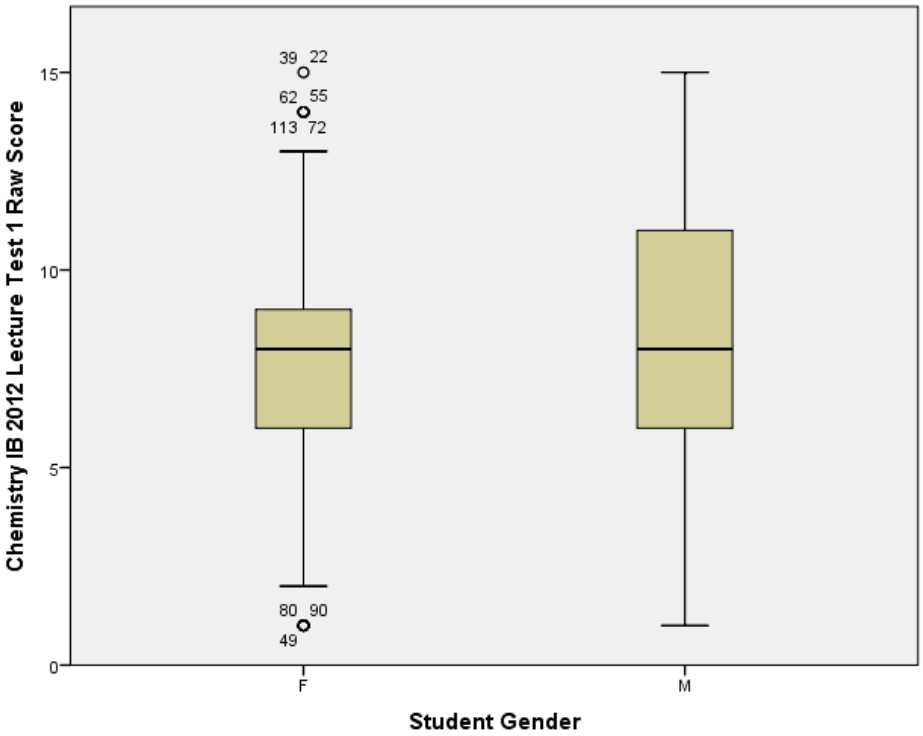

Figure 381: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Lecture Test 1 2012 to Observe Significant Differences

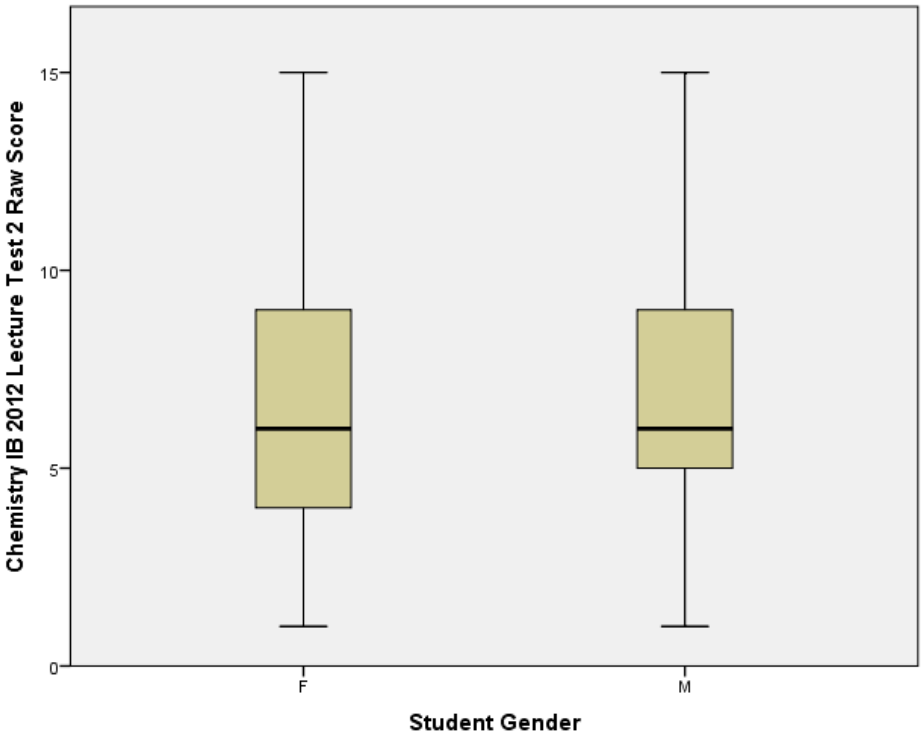

Figure 382: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Lecture Test 2 2012 to Observe Significant Differences

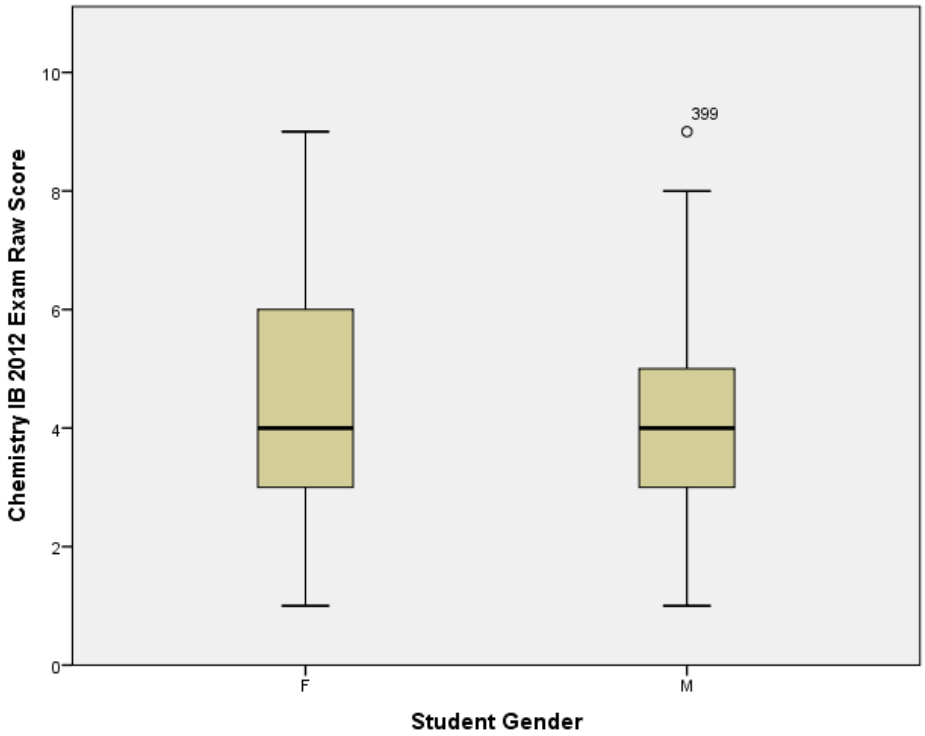

Figure 383: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Exam 2012 to Observe Significant Differences

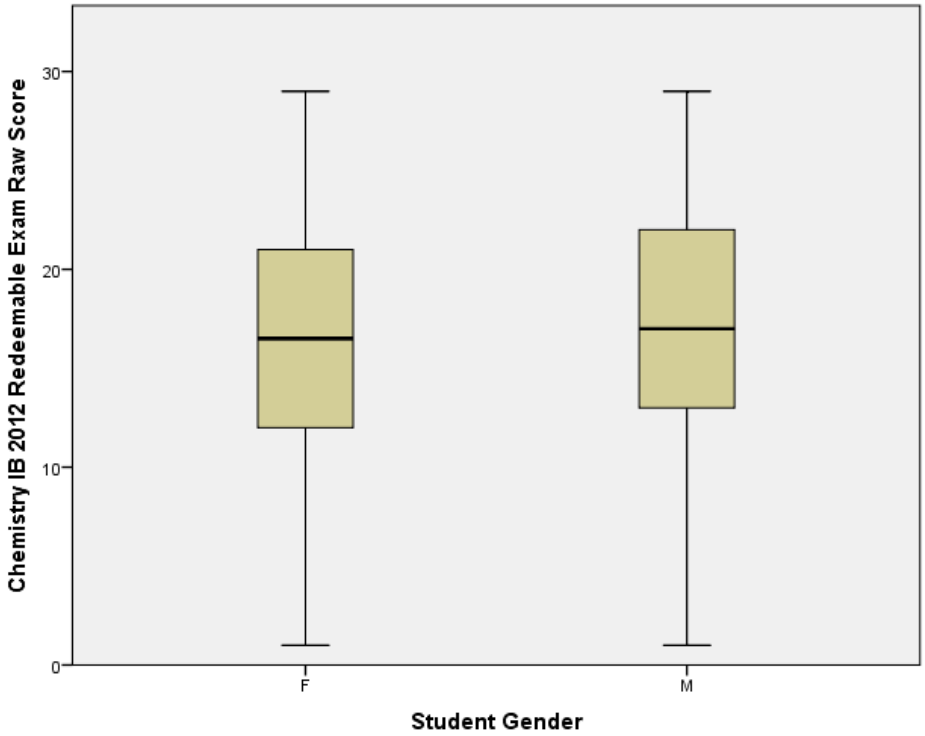

Figure 384: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Redeemable Exam 2012 to Observe Significant Differences

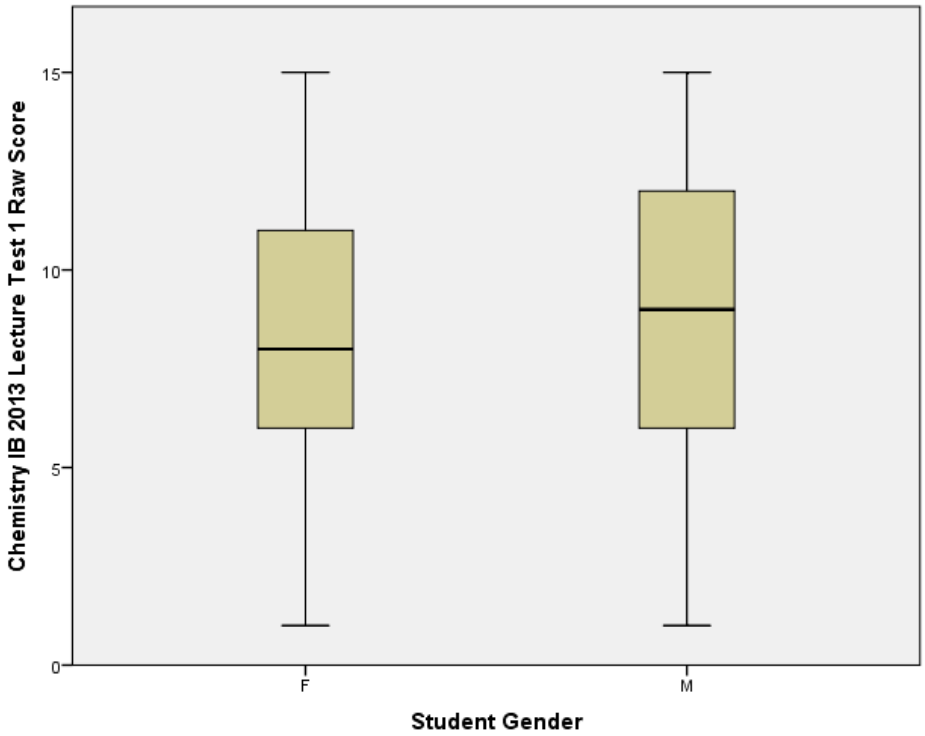

Figure 385: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Lecture Test 1 2013 to Observe Significant Differences

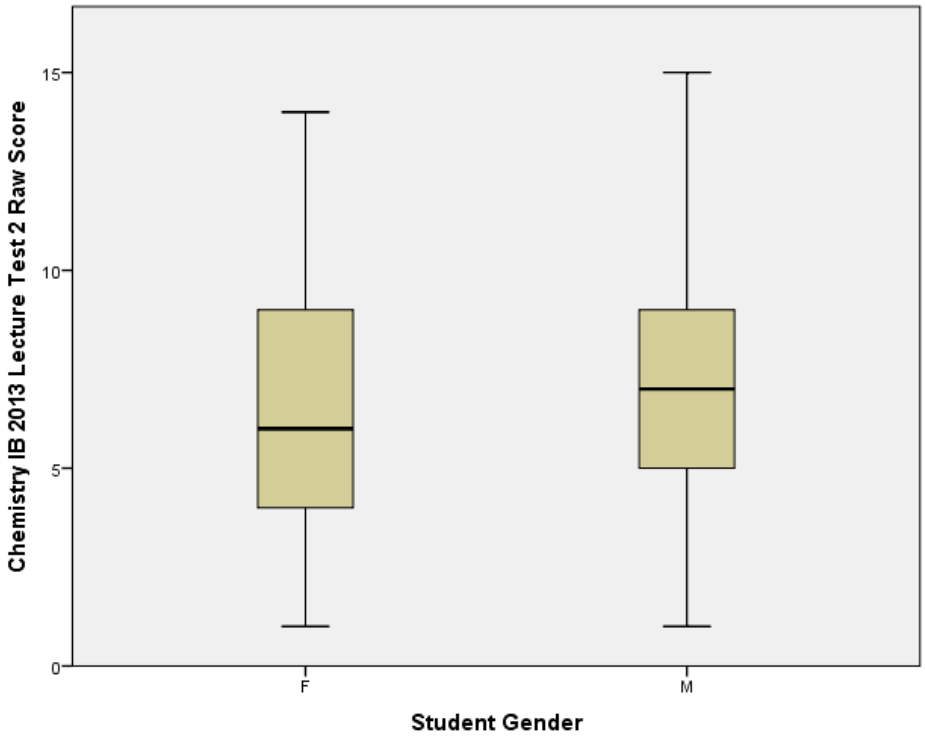

Figure 386: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Lecture Test 2 2013 to Observe Significant Differences

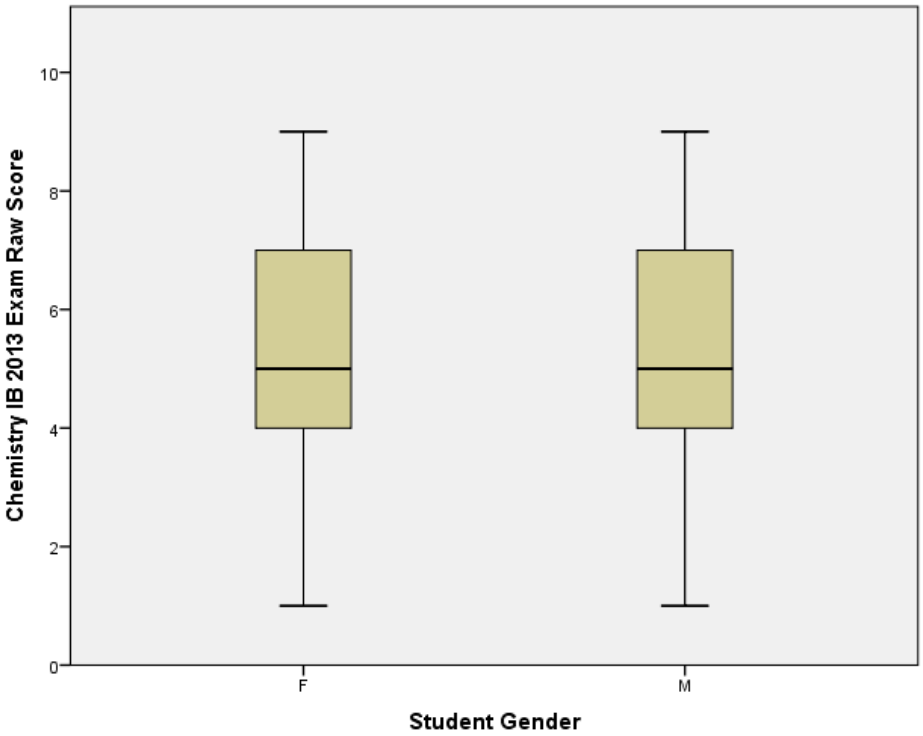

Figure 387: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Exam 2013 to Observe Significant Differences

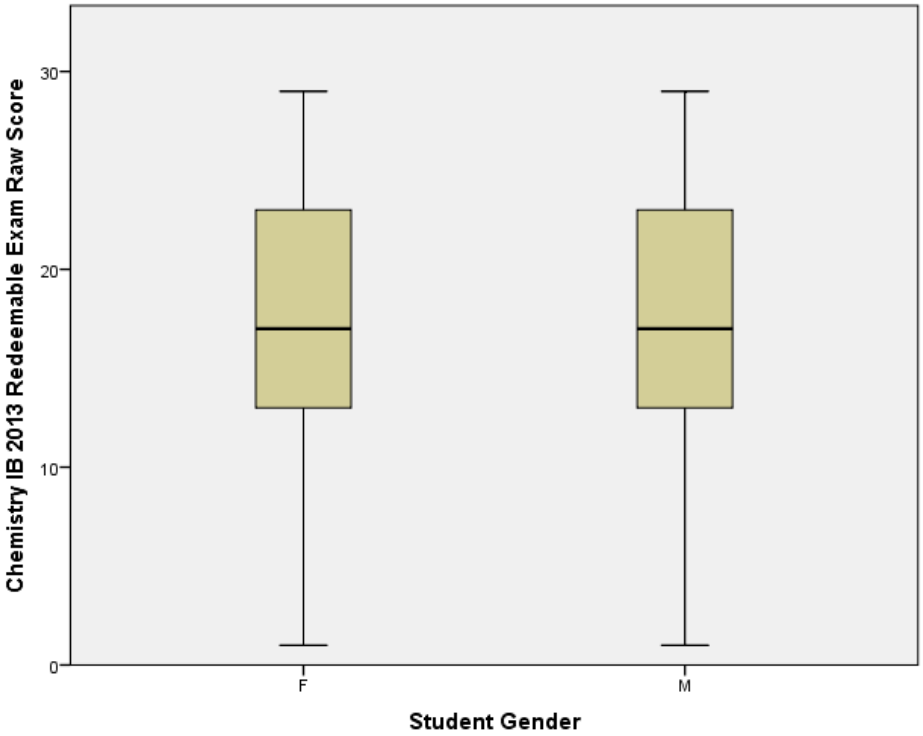

Figure 388: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Redeemable Exam 2013 to Observe Significant Differences

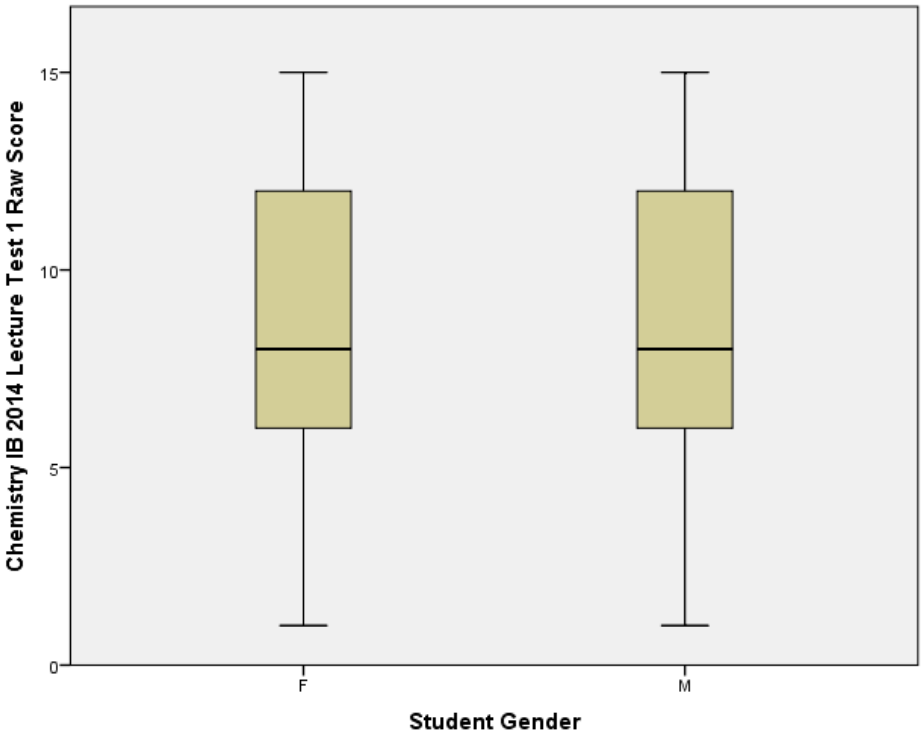

Figure 389: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Lecture Test 1 2014 to Observe Significant Differences

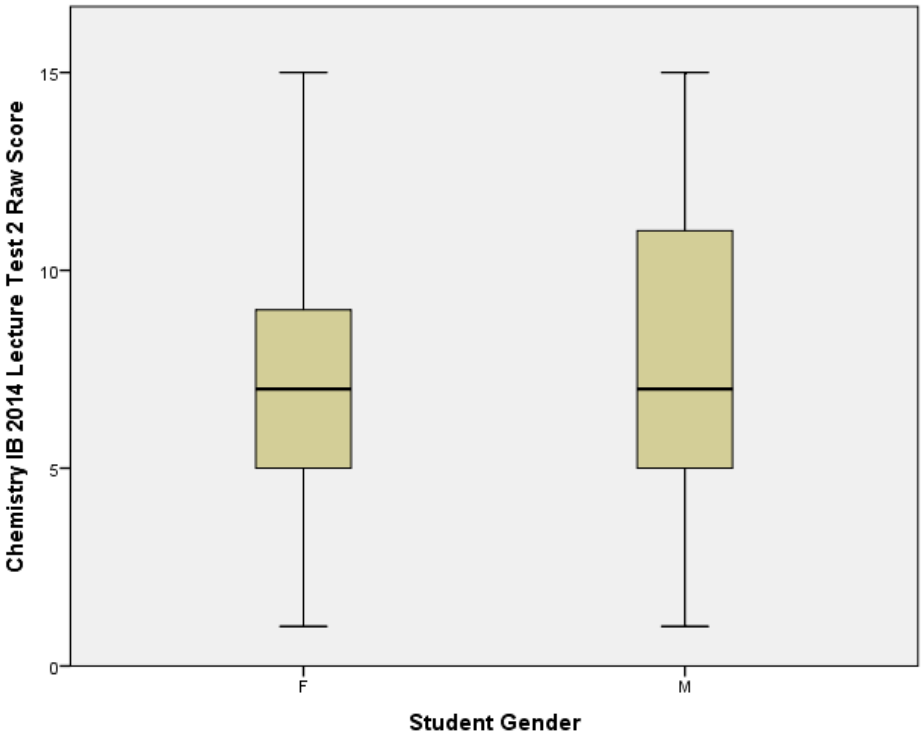

Figure 390: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Lecture Test 2 2014 to Observe Significant Differences

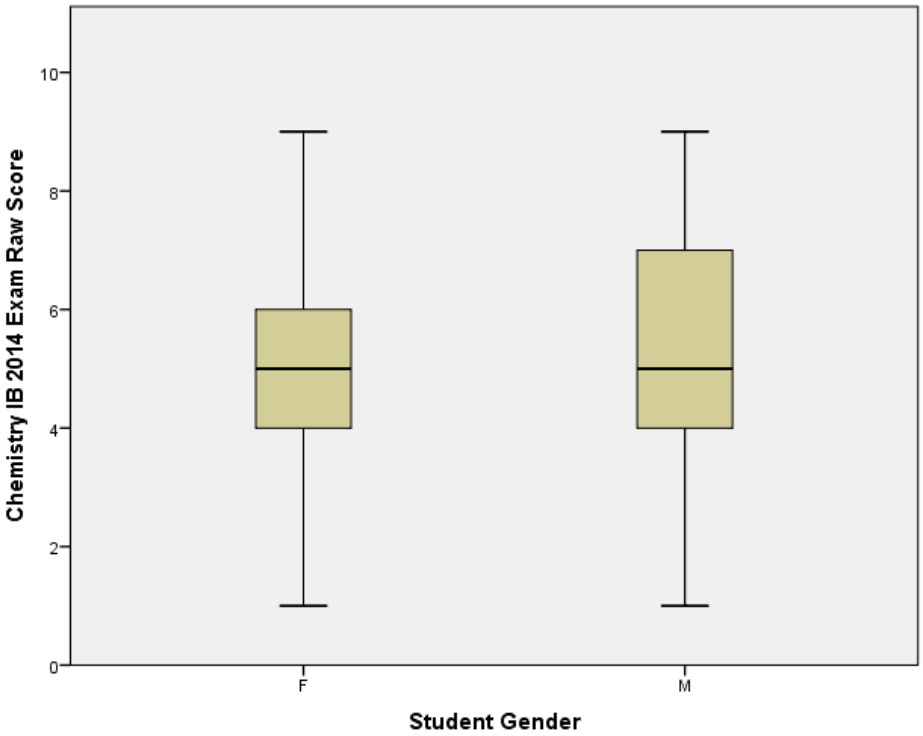

Figure 391: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Exam 2014 to Observe Significant Differences

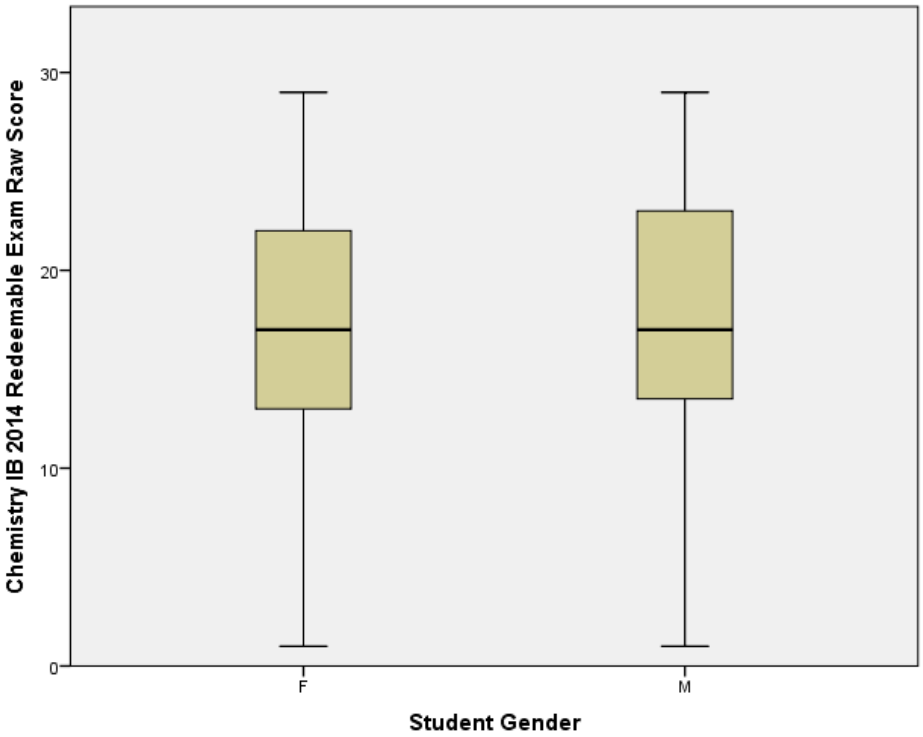

Figure 392: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Redeemable Exam 2014 to Observe Significant Differences

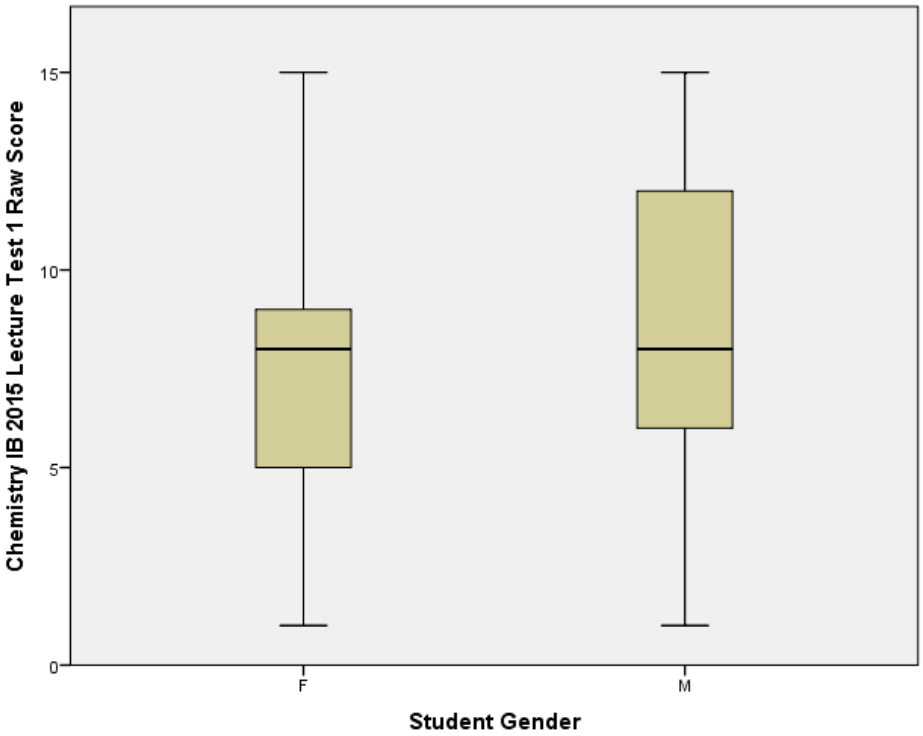

Figure 393: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Lecture Test 1 2015 to Observe Significant Differences

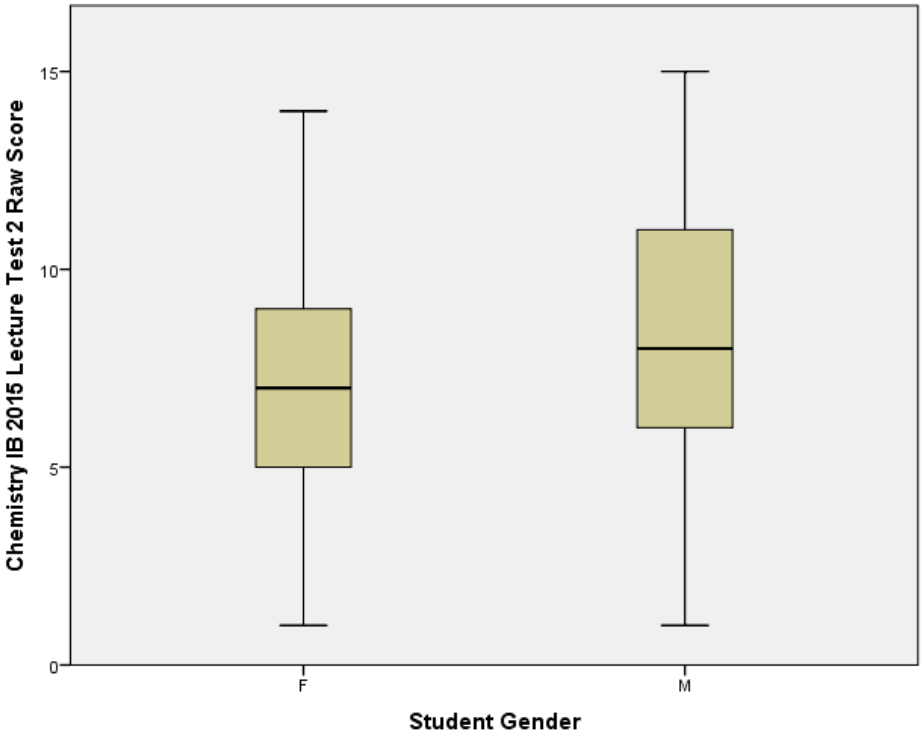

Figure 394: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Lecture Test 2 2015 to Observe Significant Differences

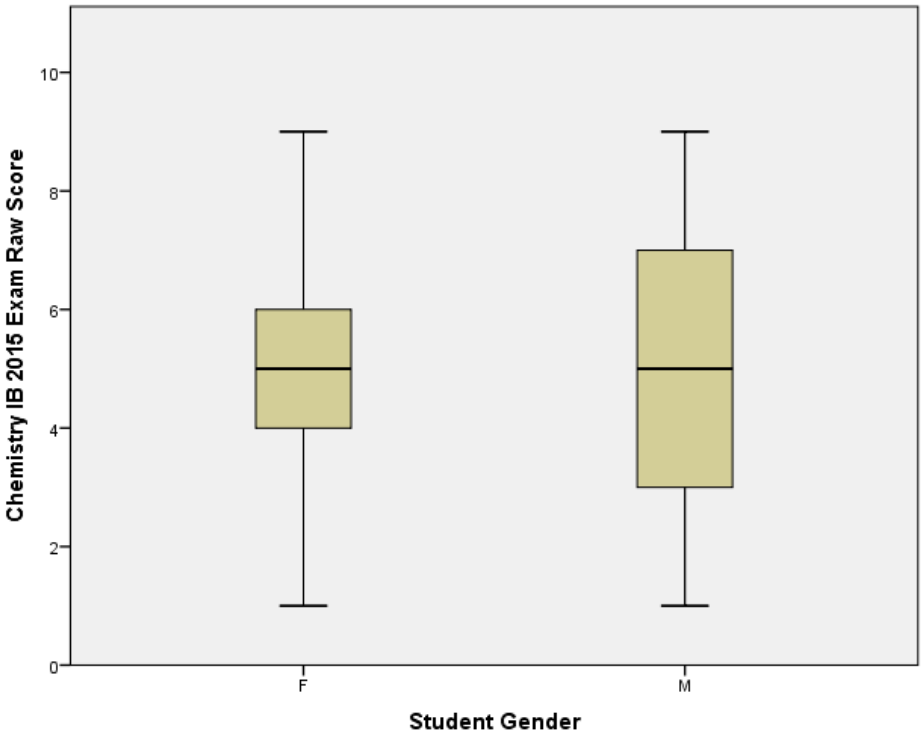

Figure 395: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Exam 2015 to Observe Significant Differences

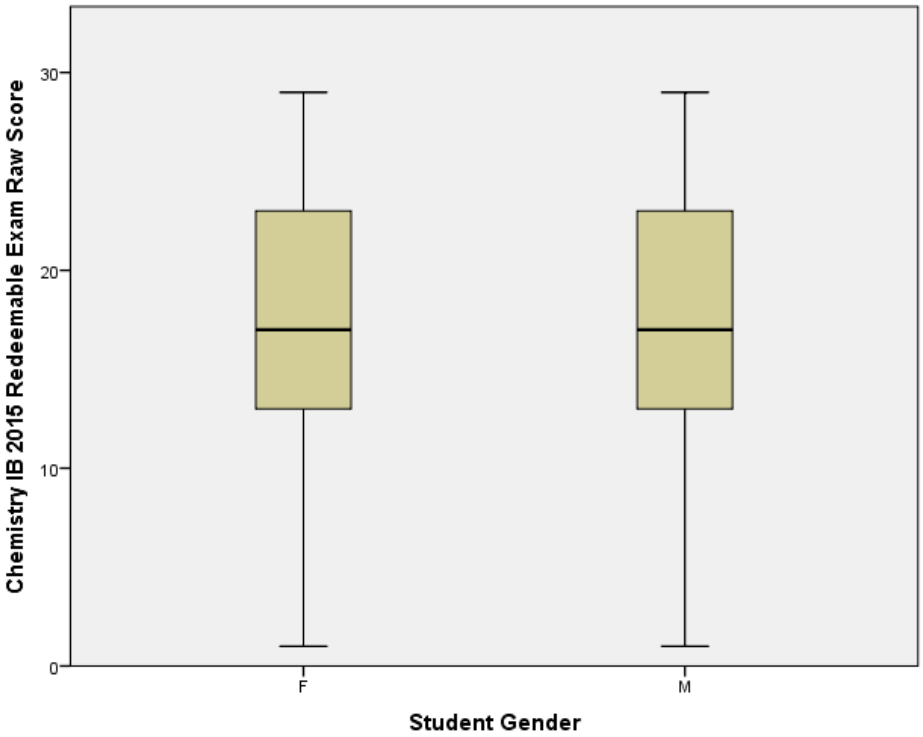

Figure 396: The Boxplot Comparison of Male and Female Raw Score in Chemistry IB Redeemable Exam 2015 to Observe Significant Differences

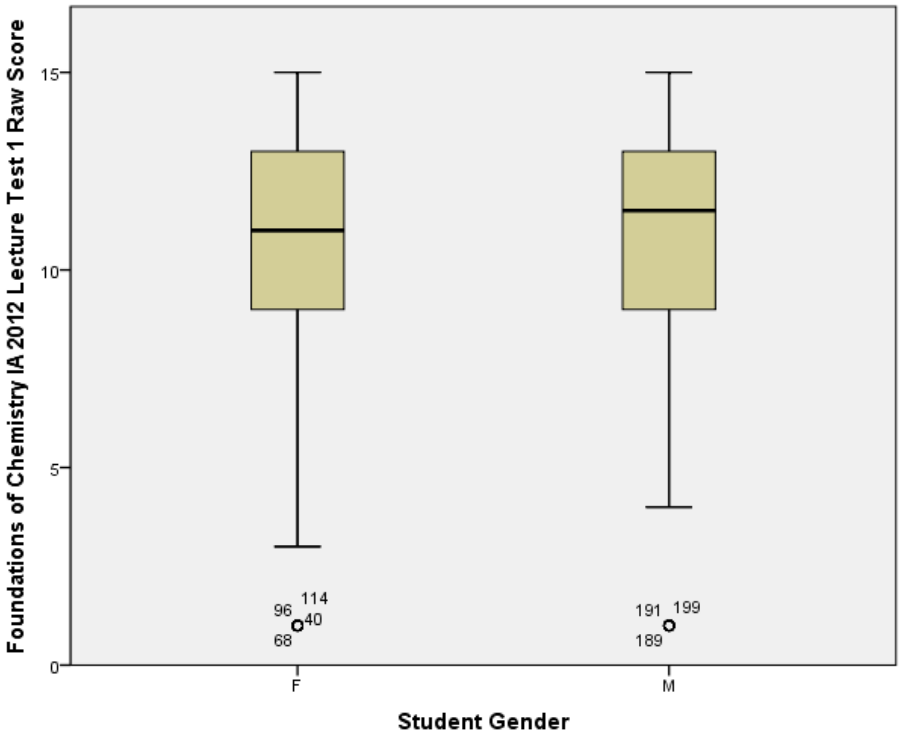

Figure 397: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Lecture Test 1 2012 to Observe Significant Differences

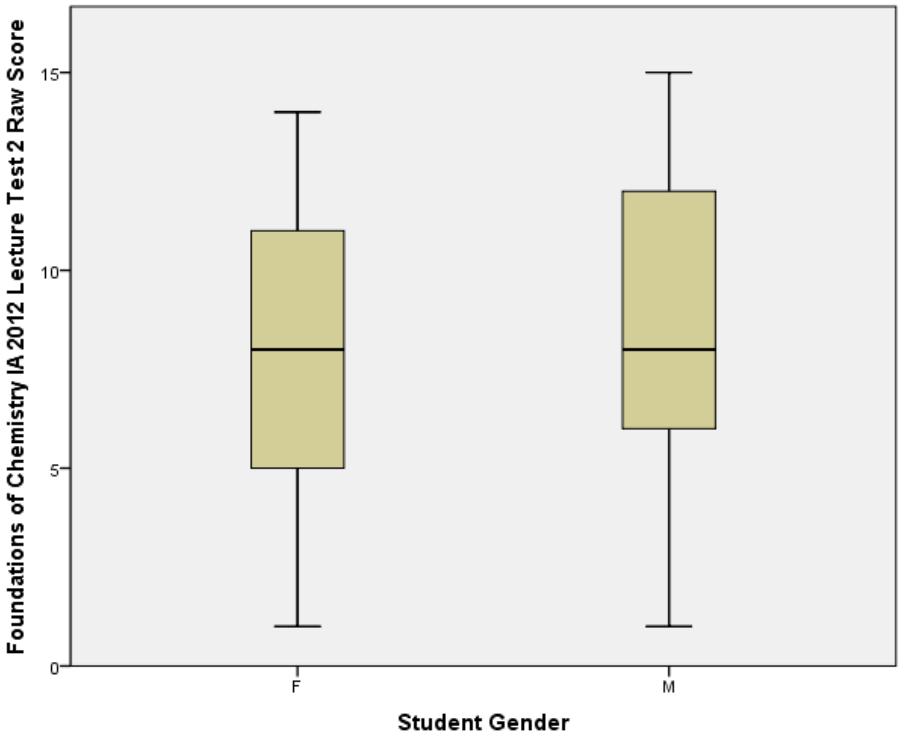

Figure 398: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Lecture Test 2 2012 to Observe Significant Differences

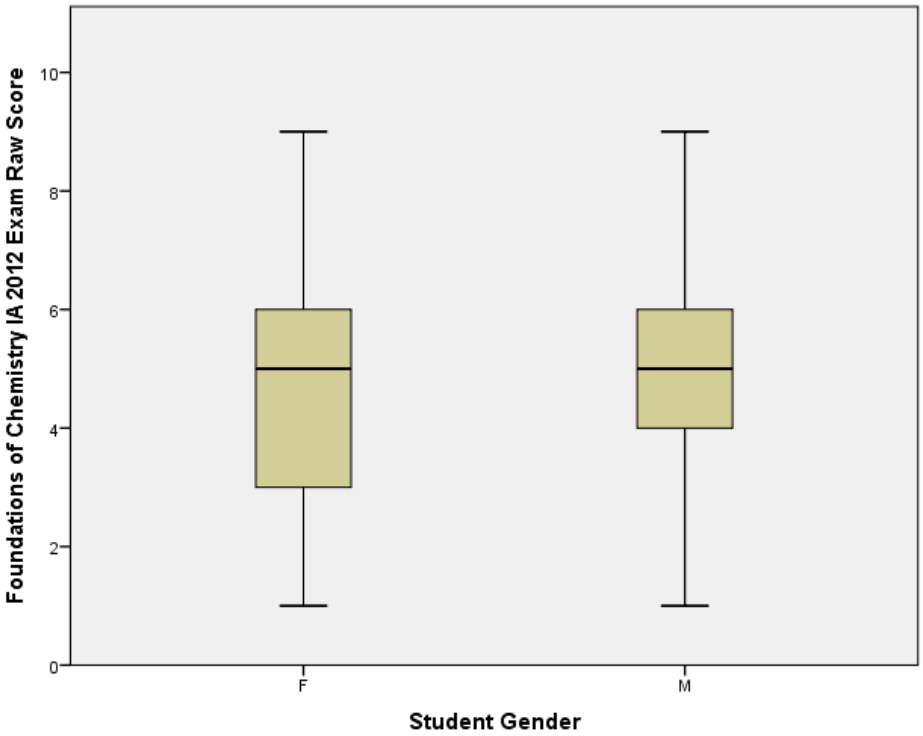

Figure 399: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Exam 2012 to Observe Significant Differences

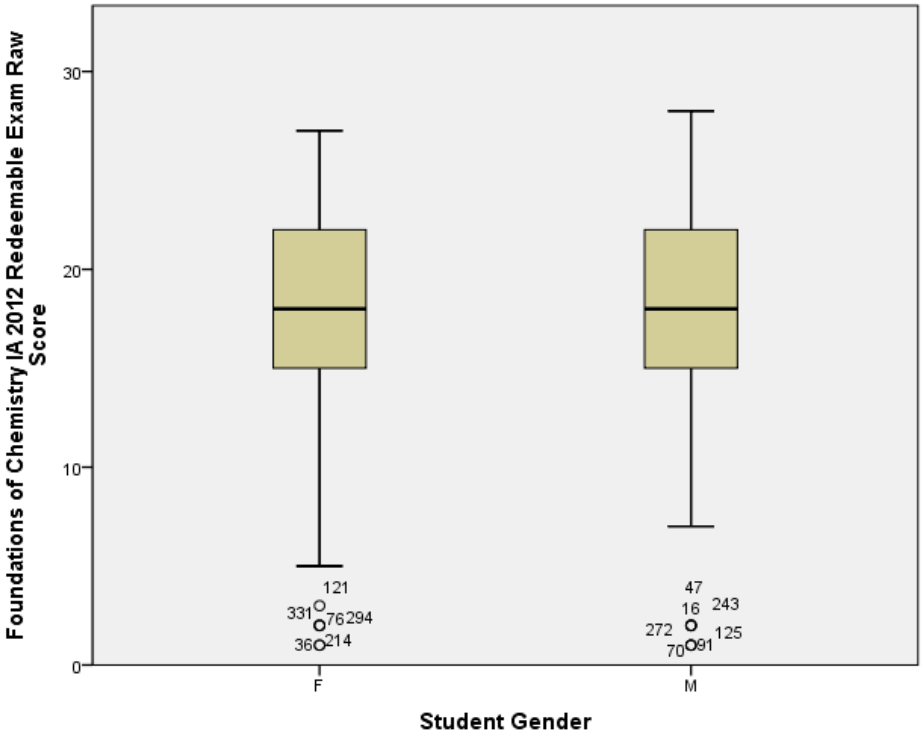

Figure 400: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Redeemable Exam 2012 to Observe Significant Differences

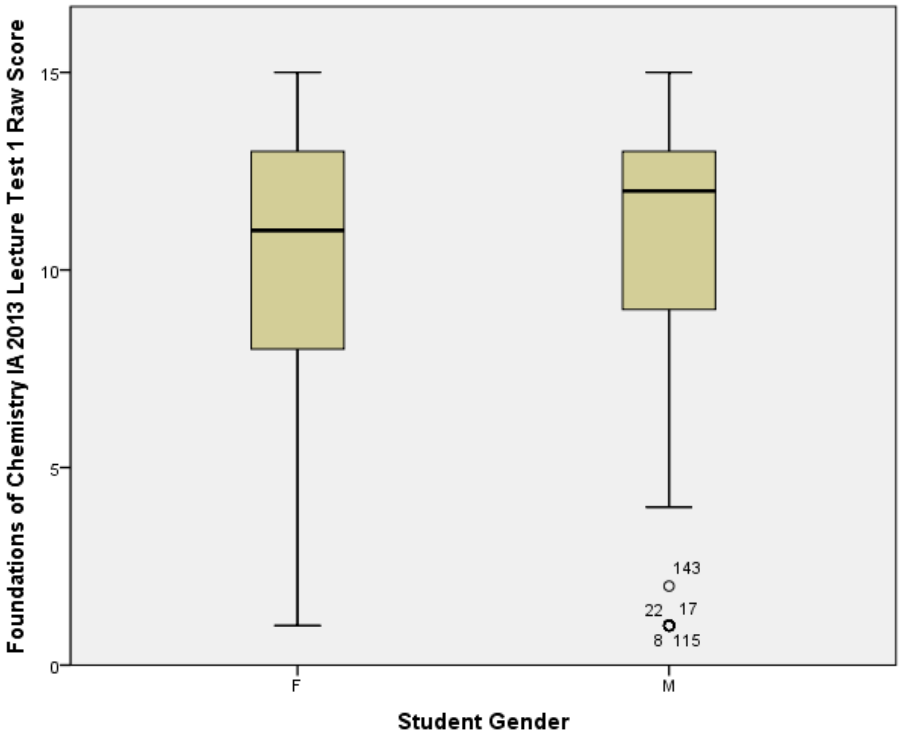

Figure 401: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Lecture Test 1 2013 to Observe Significant Differences

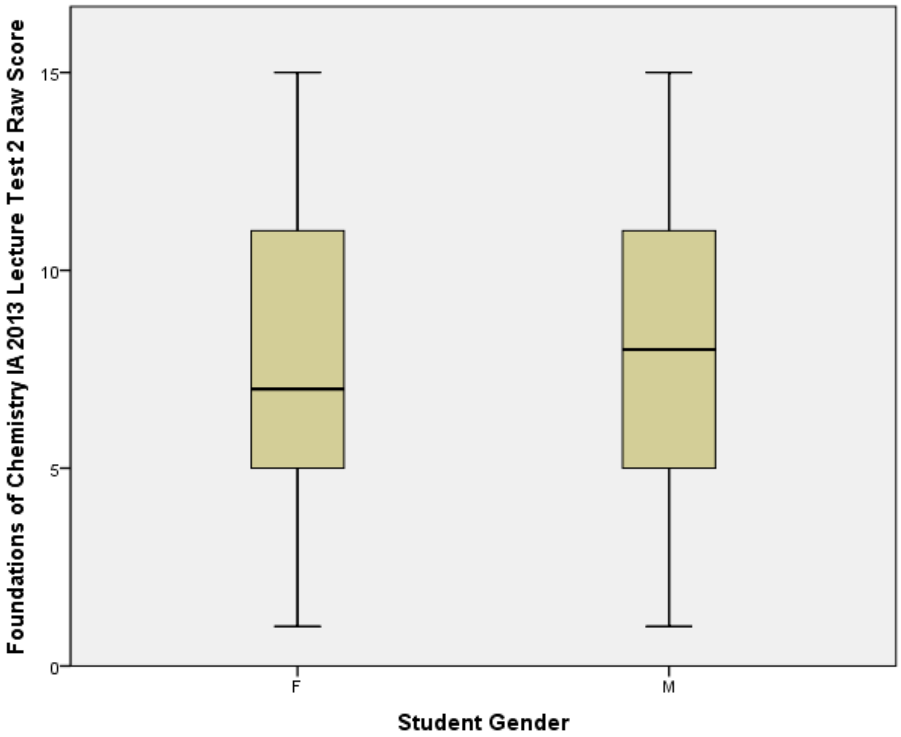

Figure 402: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Lecture Test 2 2013 to Observe Significant Differences

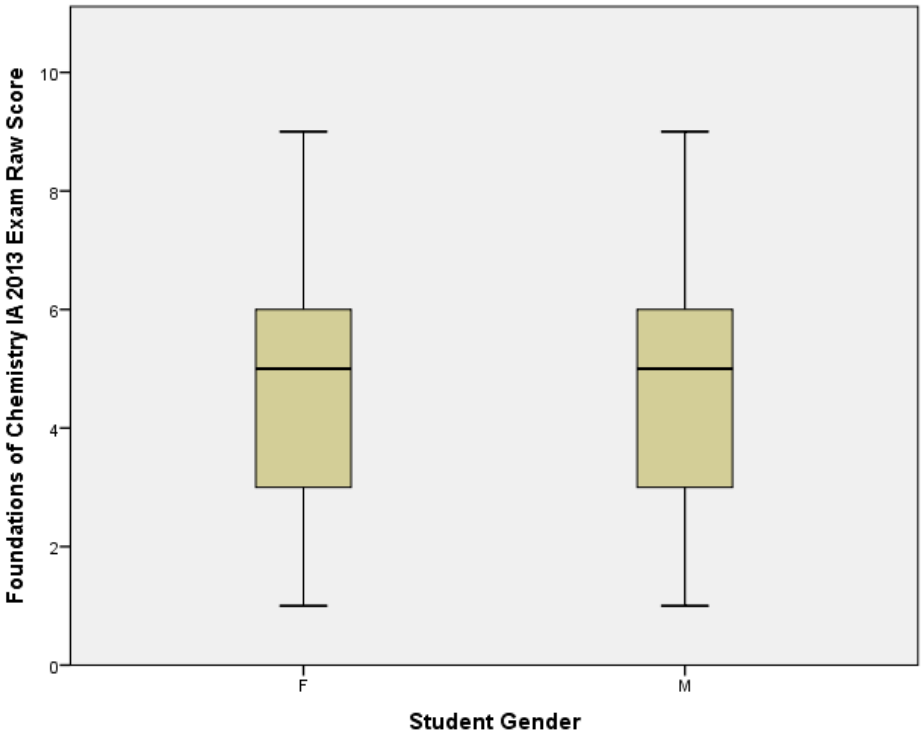

Figure 403: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Exam 2013 to Observe Significant Differences

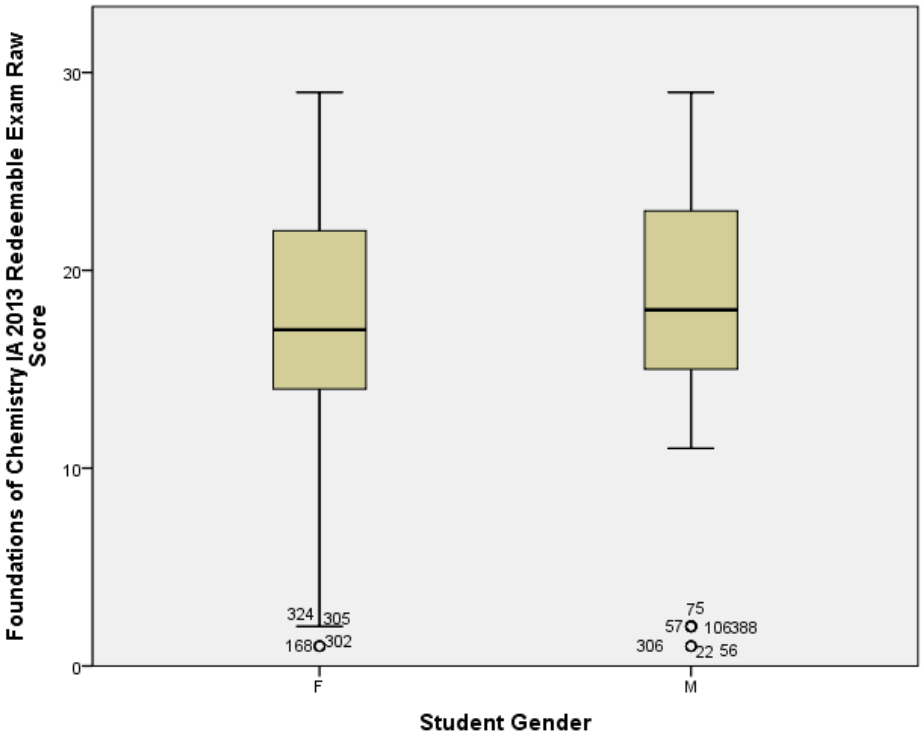

Figure 404: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Redeemable Exam 2013 to Observe Significant Differences

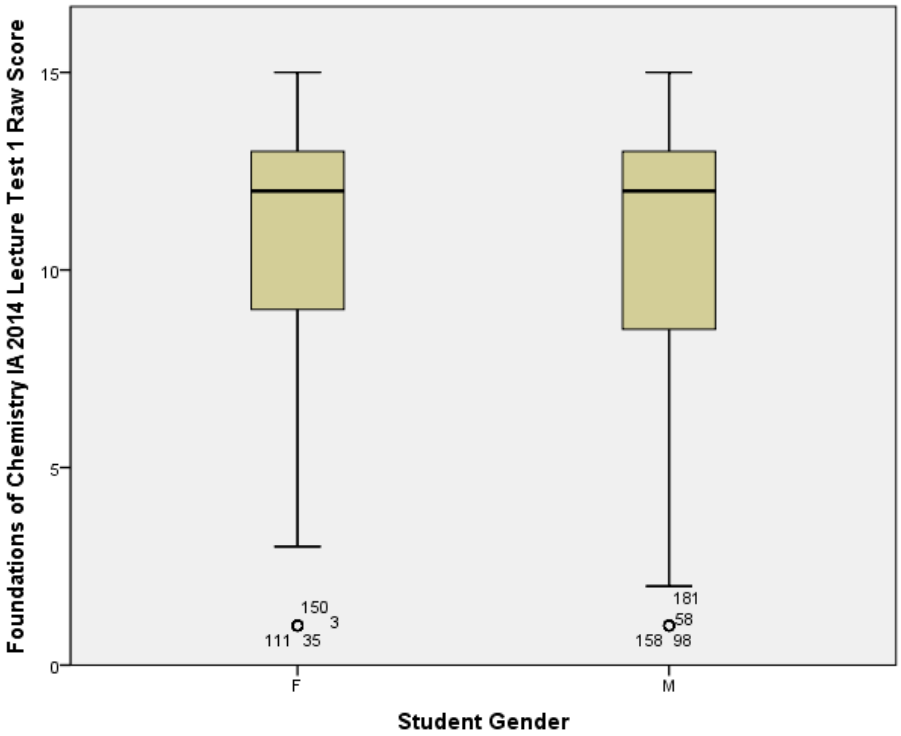

Figure 405: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Lecture Test 1 2014 to Observe Significant Differences

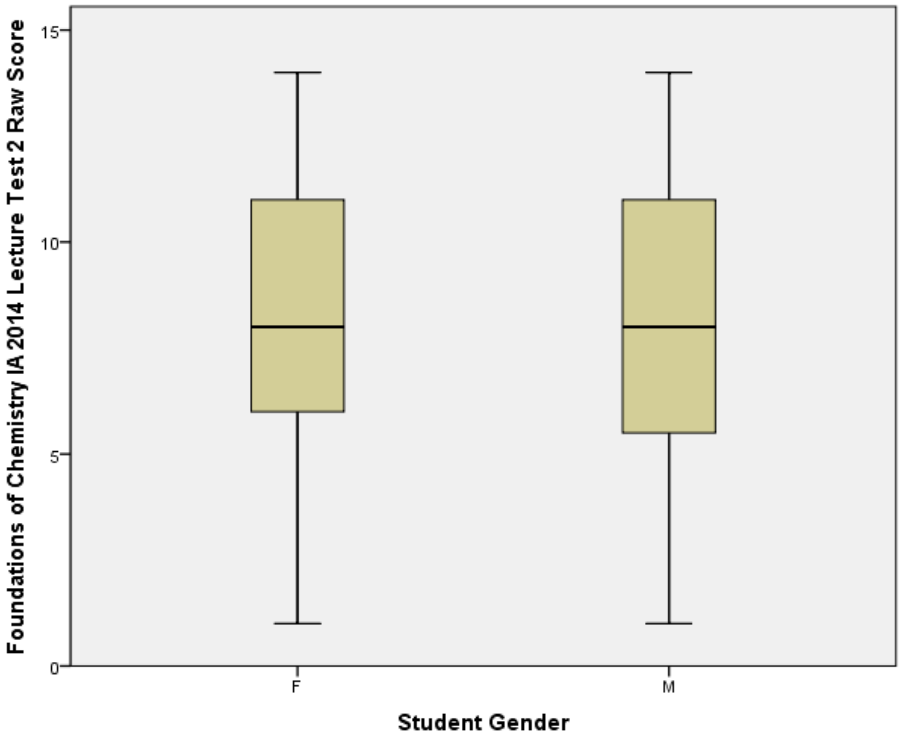

Figure 406: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Lecture Test 2 2014 to Observe Significant Differences

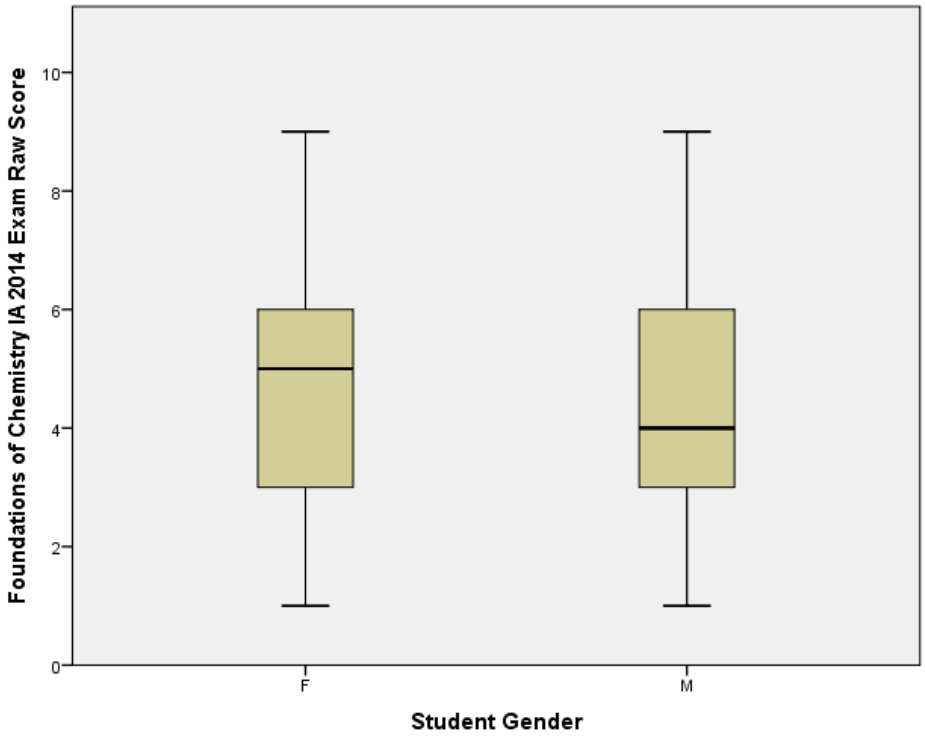

Figure 407: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Exam 2014 to Observe Significant Differences

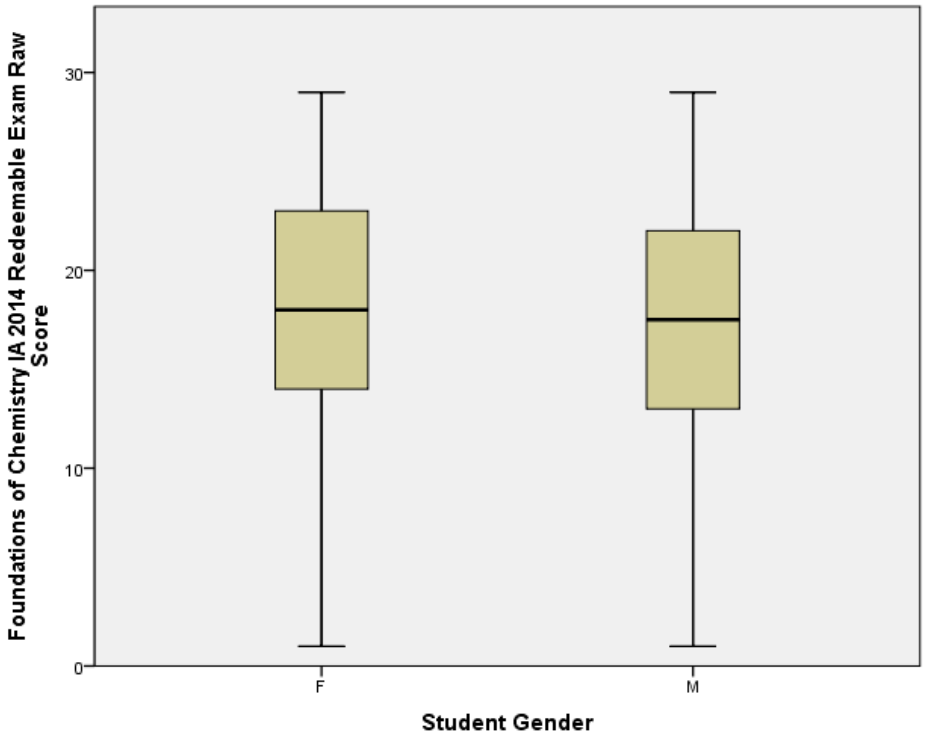

Figure 408: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Redeemable Exam 2014 to Observe Significant Differences

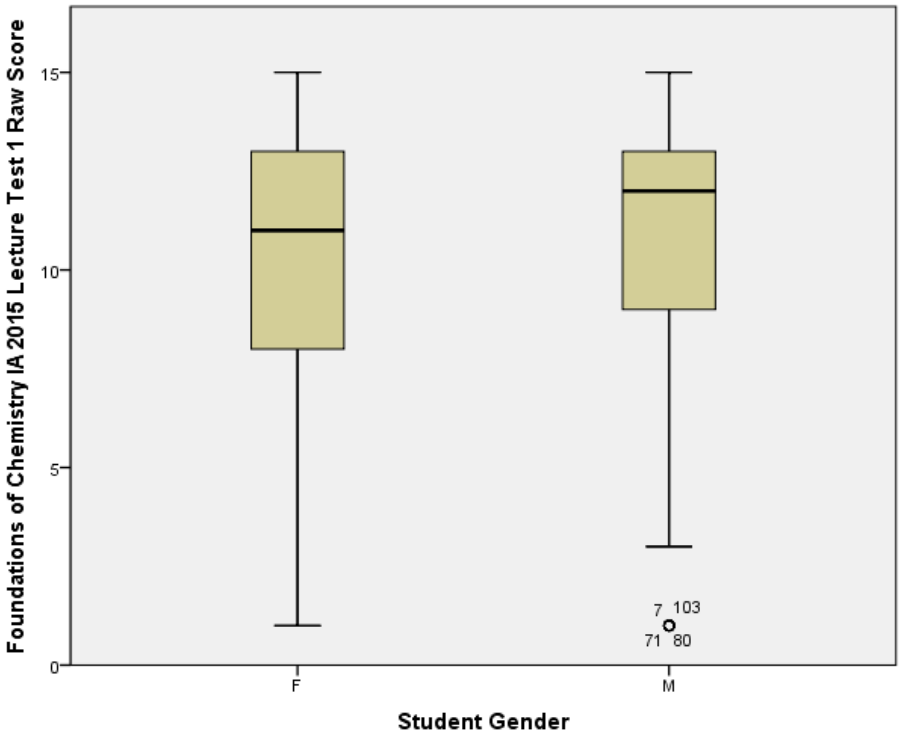

Figure 409: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Lecture Test 1 2015 to Observe Significant Differences

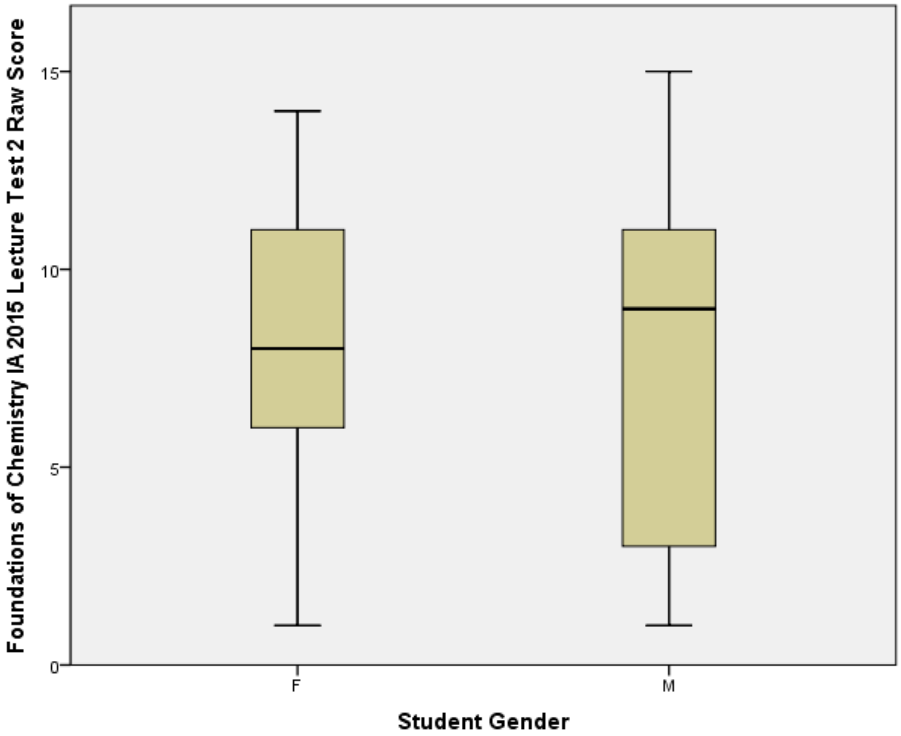

Figure 410: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Lecture Test 2 2015 to Observe Significant Differences

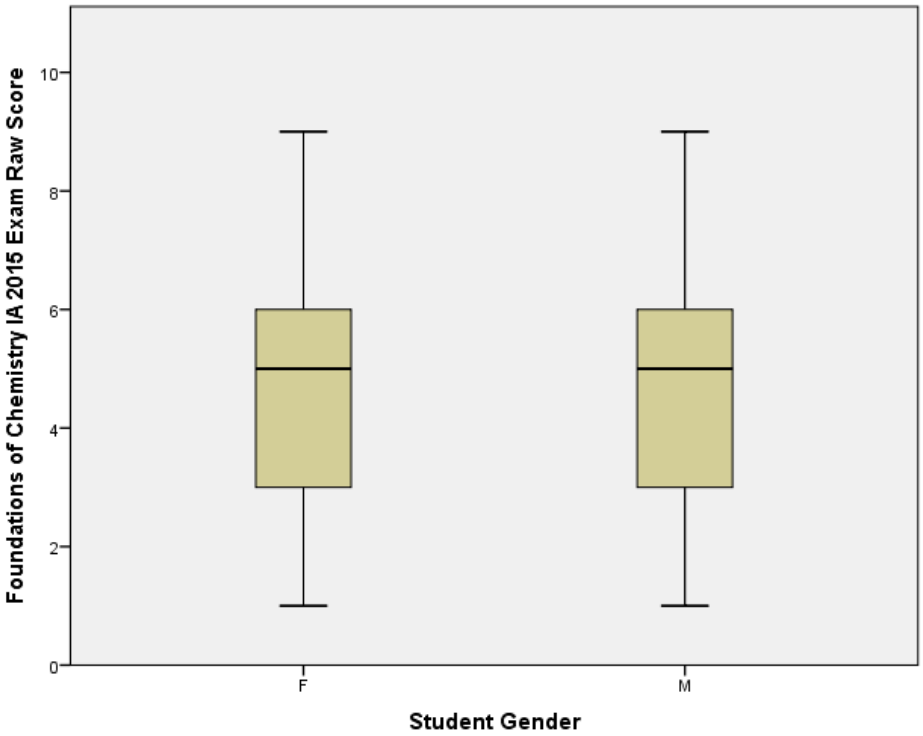

Figure 411: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Exam 2015 to Observe Significant Differences

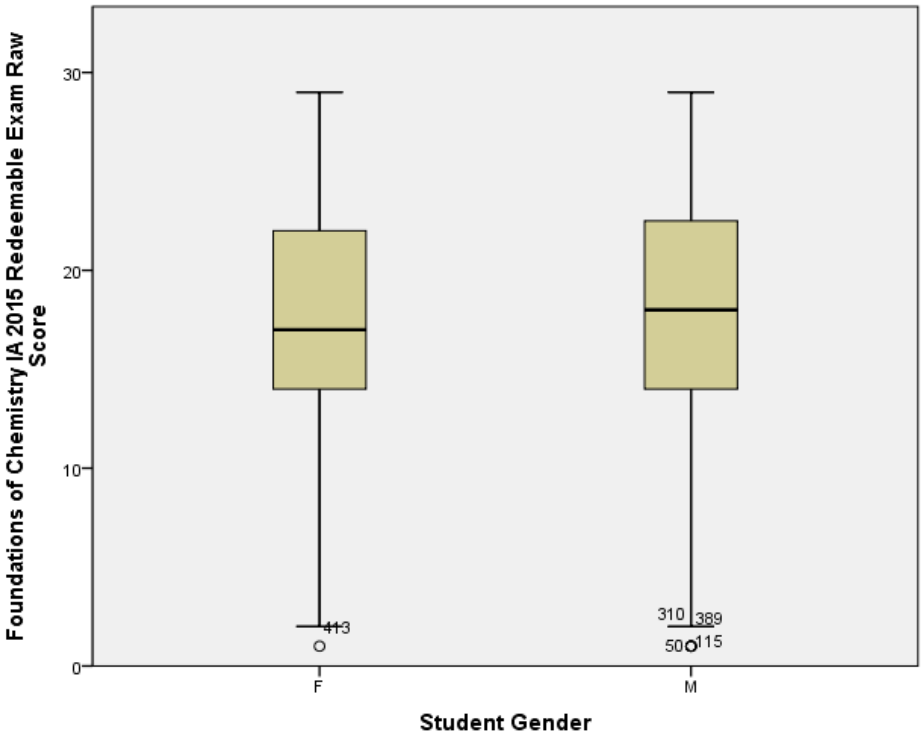

Figure 412: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IA Redeemable Exam 2015 to Observe Significant Differences

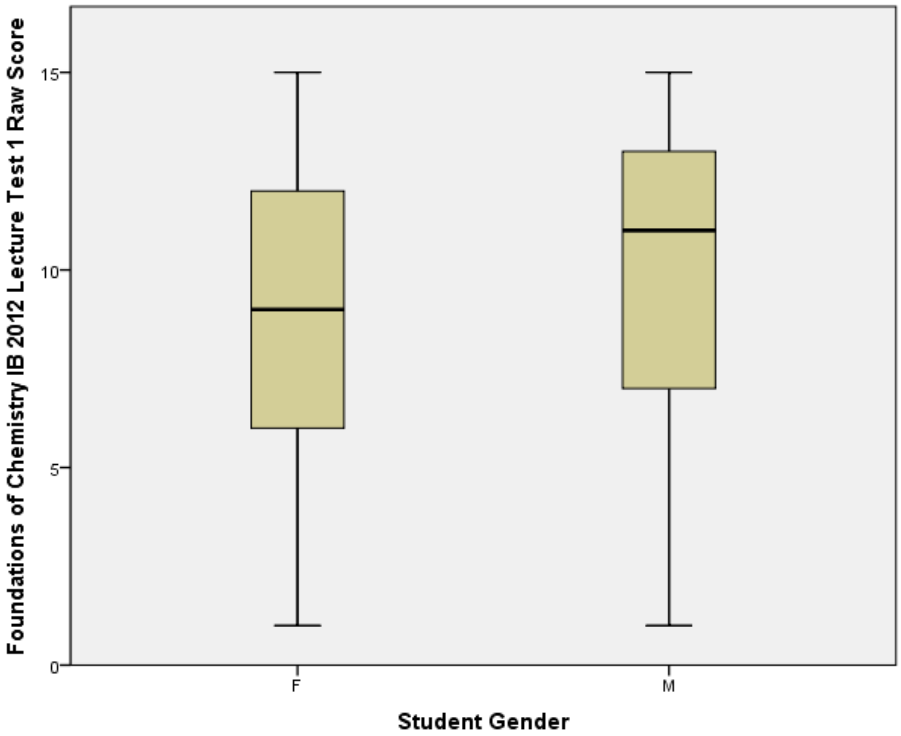

Figure 413: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Lecture Test 1 2012 to Observe Significant Differences

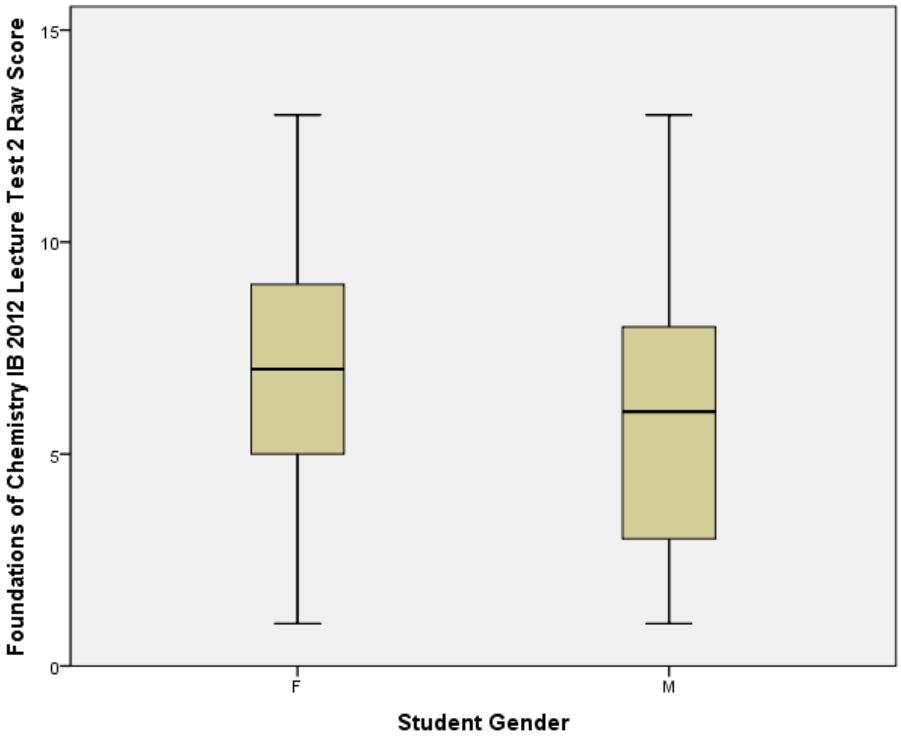

Figure 414: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Lecture Test 2 2012 to Observe Significant Differences

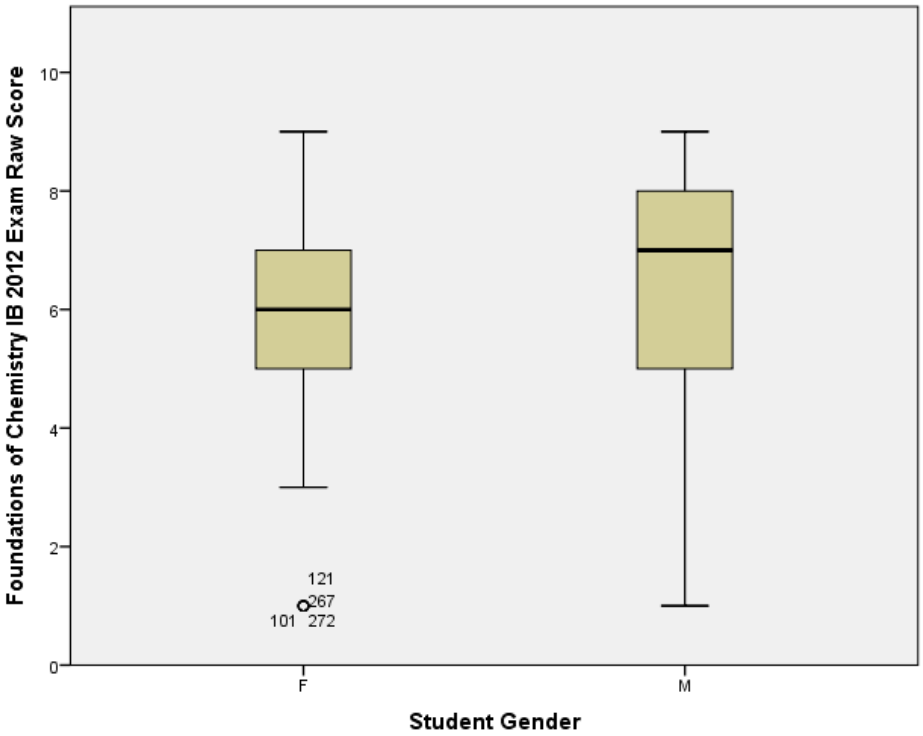

Figure 415: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Exam 2012 to Observe Significant Differences

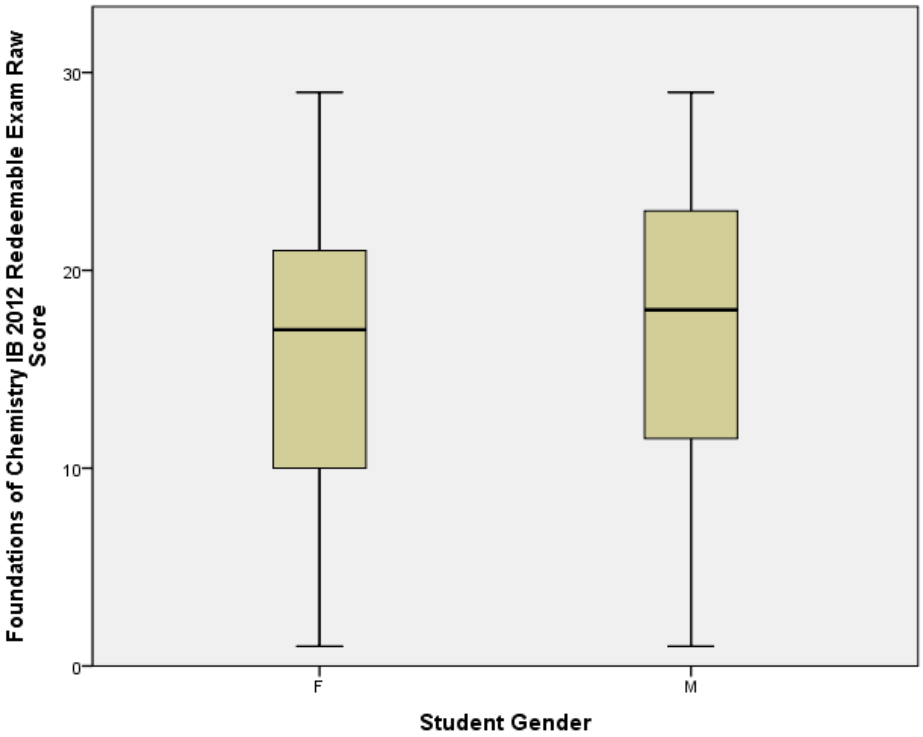

Figure 416: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Redeemable Exam 2012 to Observe Significant Differences

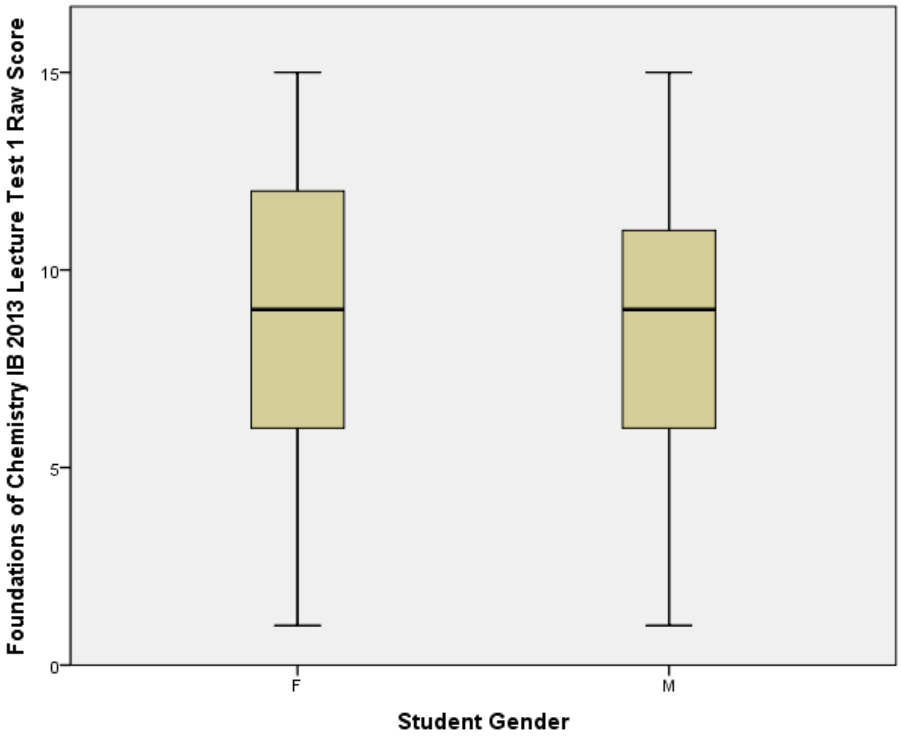

Figure 417: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Lecture Test 1 2013 to Observe Significant Differences

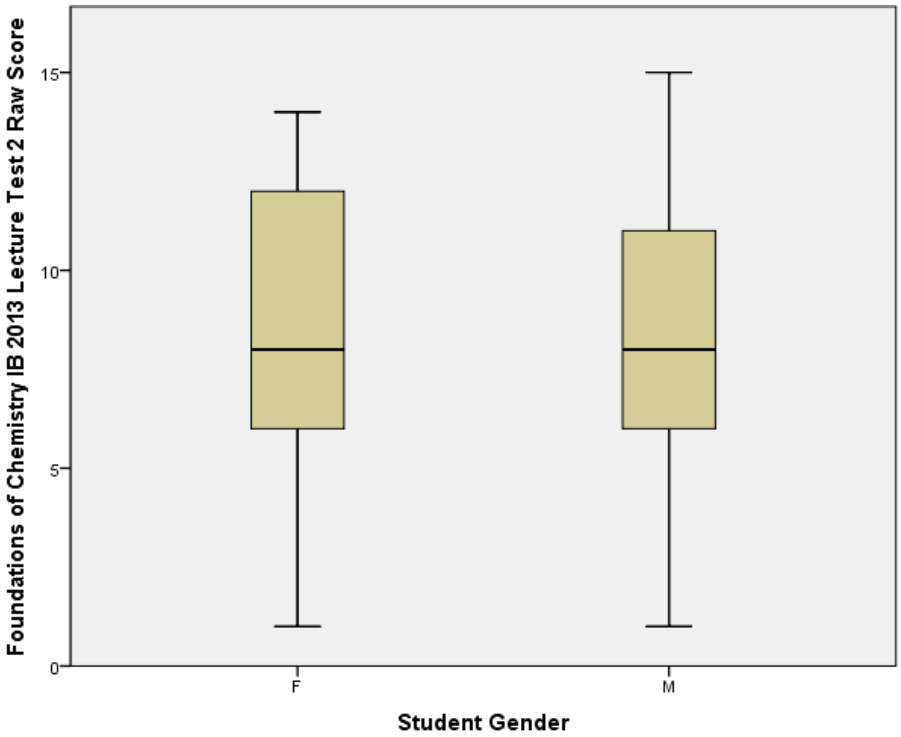

Figure 418: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Lecture Test 2 2013 to Observe Significant Differences

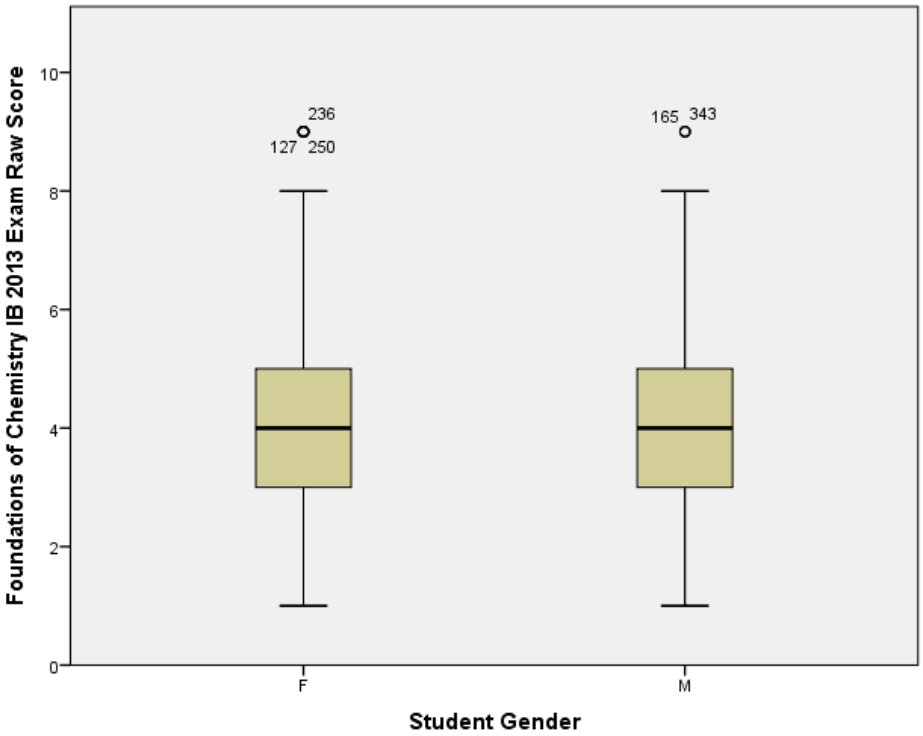

Figure 419: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Exam 2013 to Observe Significant Differences

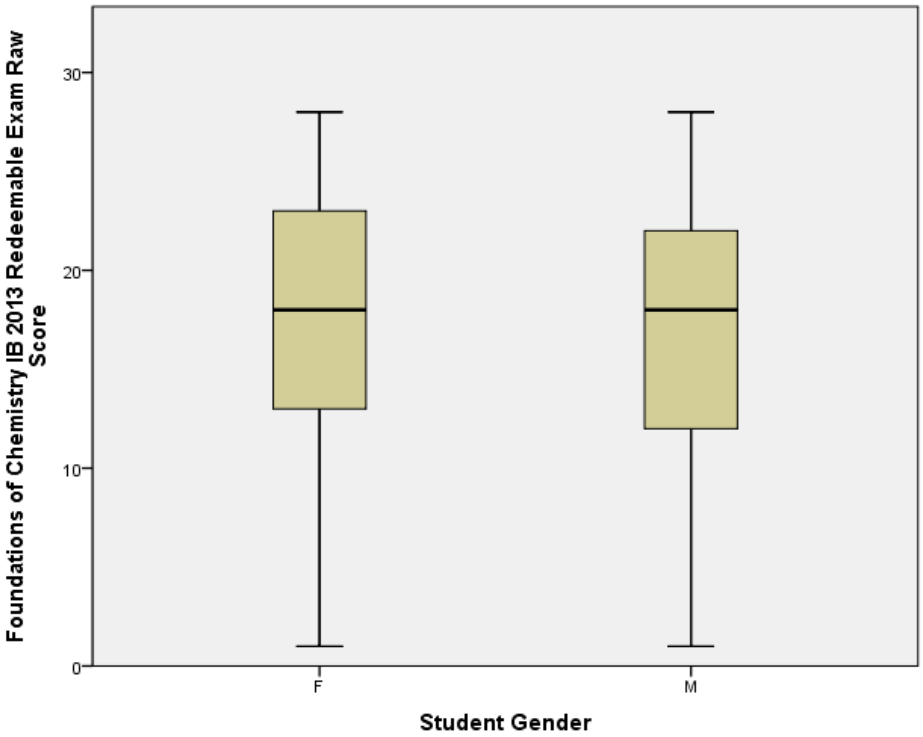

Figure 420: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Redeemable Exam 2013 to Observe Significant Differences

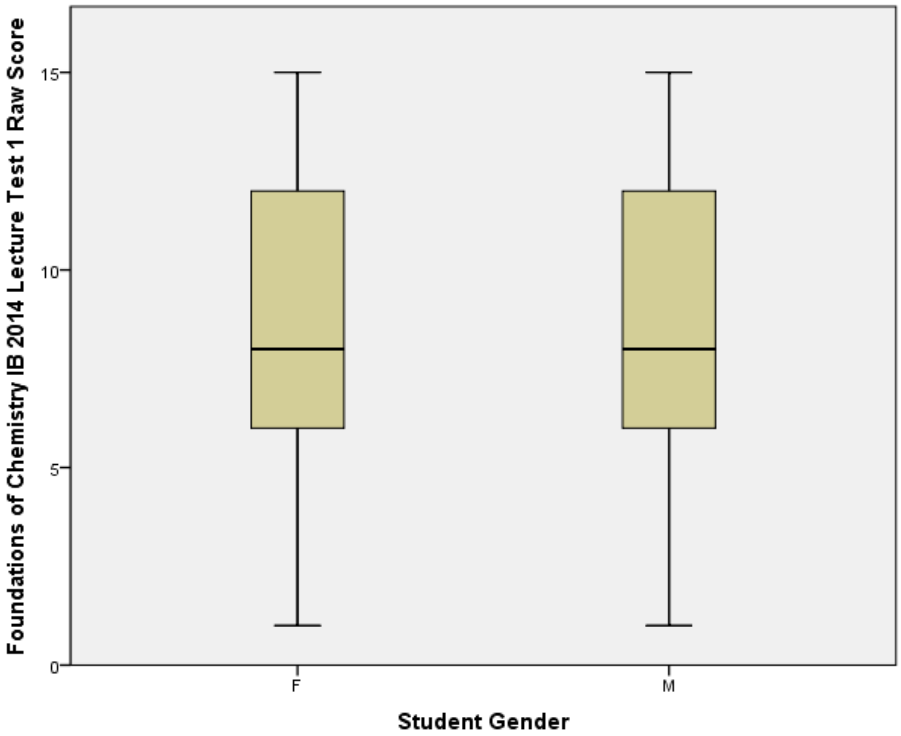

Figure 421: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Lecture Test 1 2014 to Observe Significant Differences

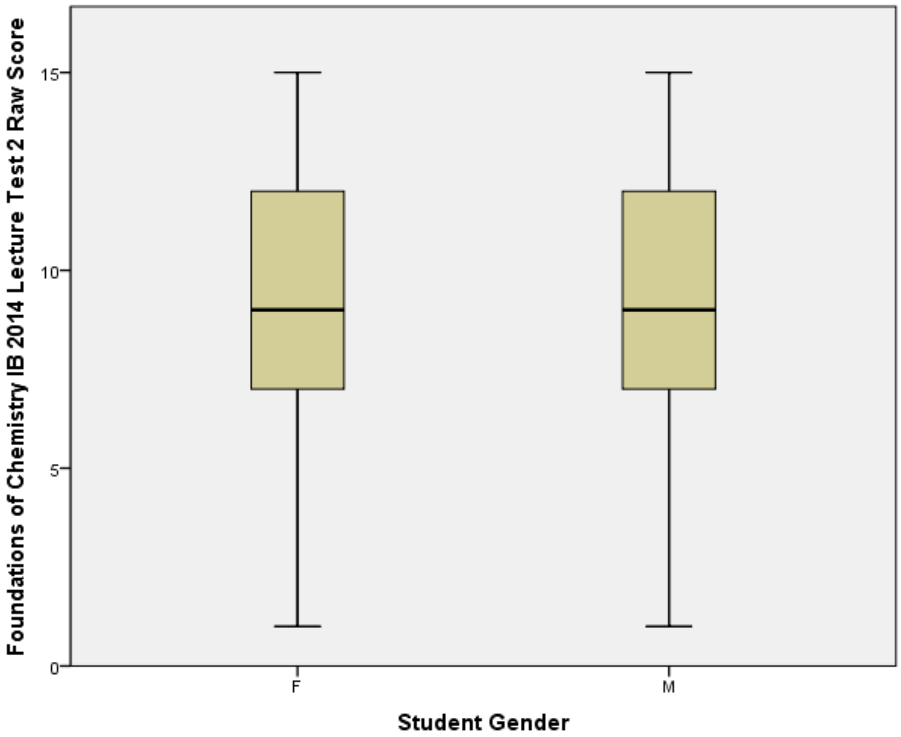

Figure 422: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Lecture Test 2 2014 to Observe Significant Differences

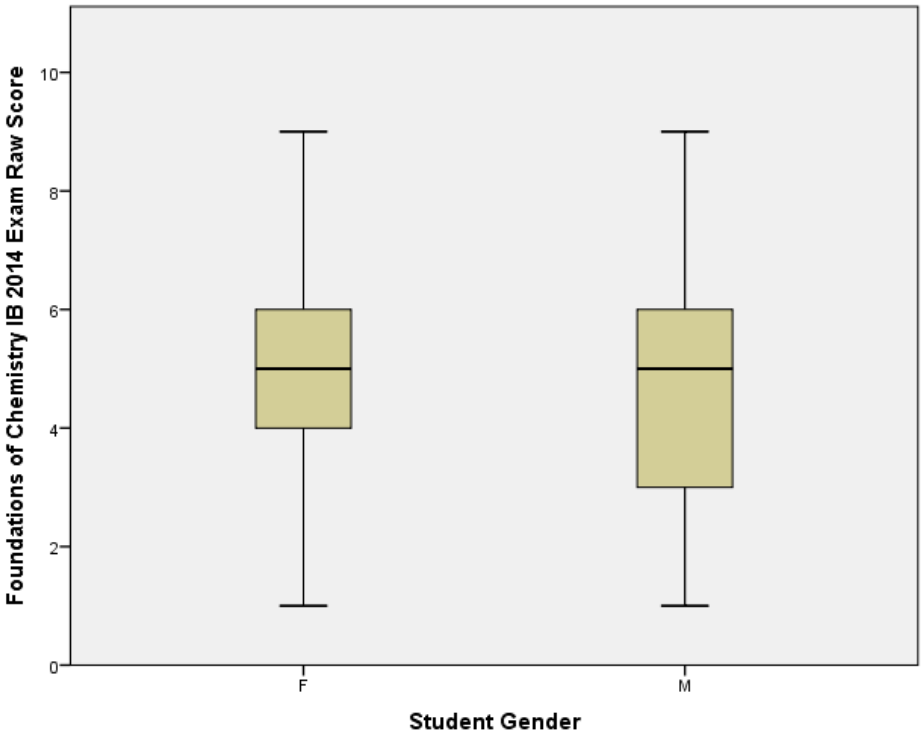

Figure 423: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Exam 2014 to Observe Significant Differences

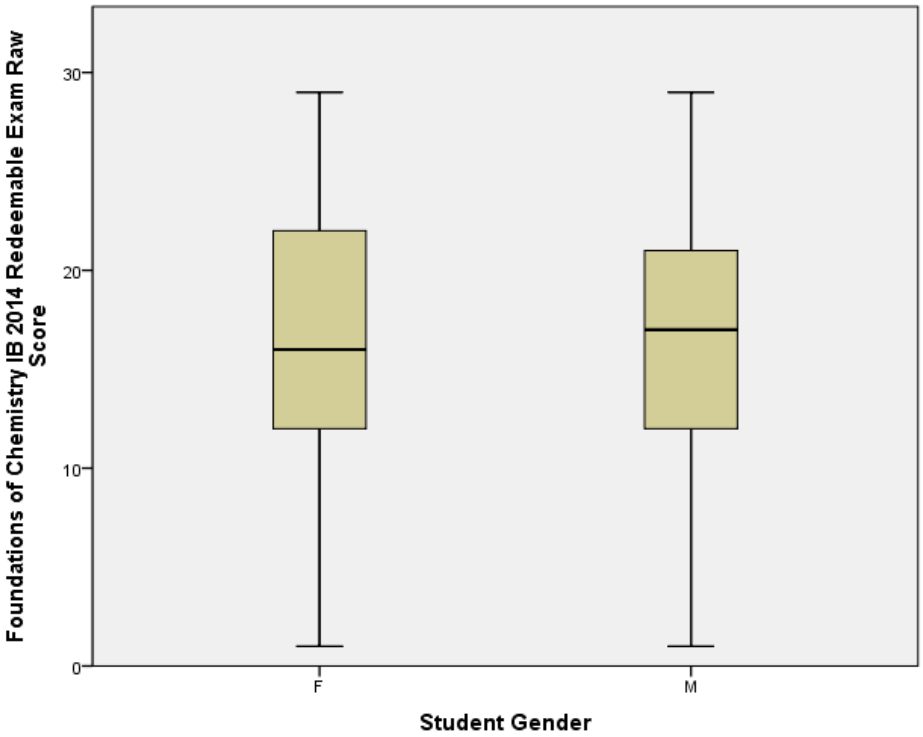

Figure 424: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Redeemable Exam 2014 to Observe Significant Differences

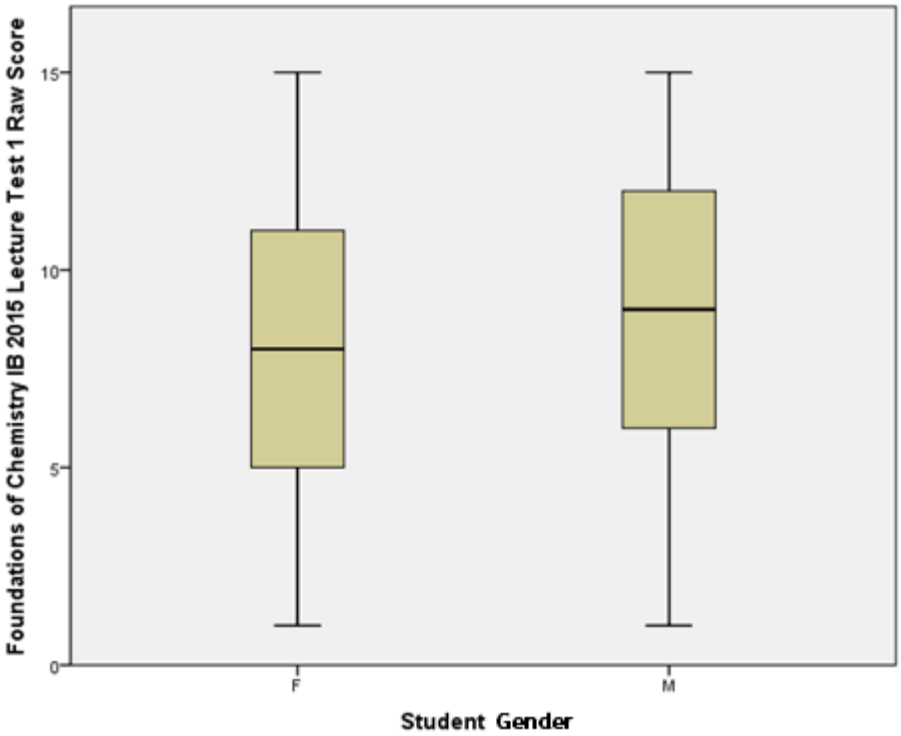

Figure 425: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Lecture Test 1 2015 to Observe Significant Differences

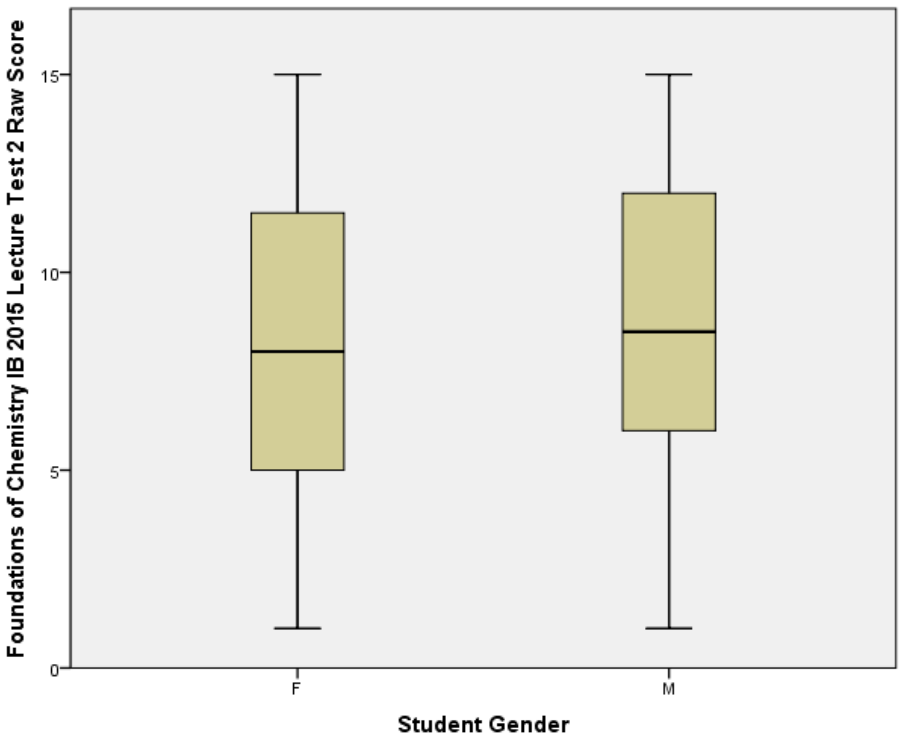

Figure 426: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Lecture Test 2 2015 to Observe Significant Differences

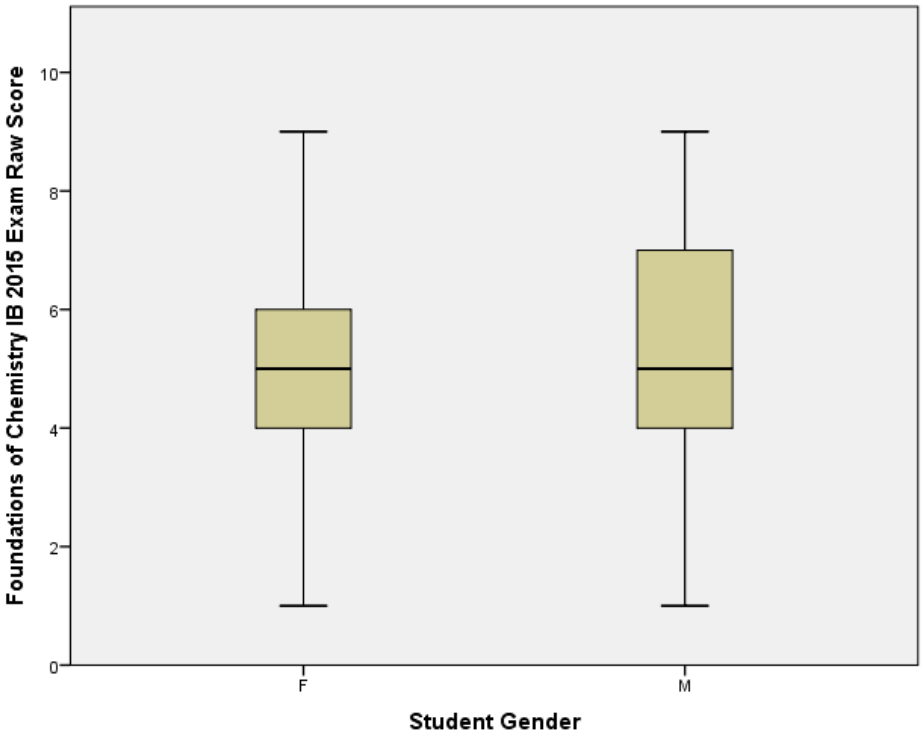

Figure 427: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Exam 2015 to Observe Significant Differences

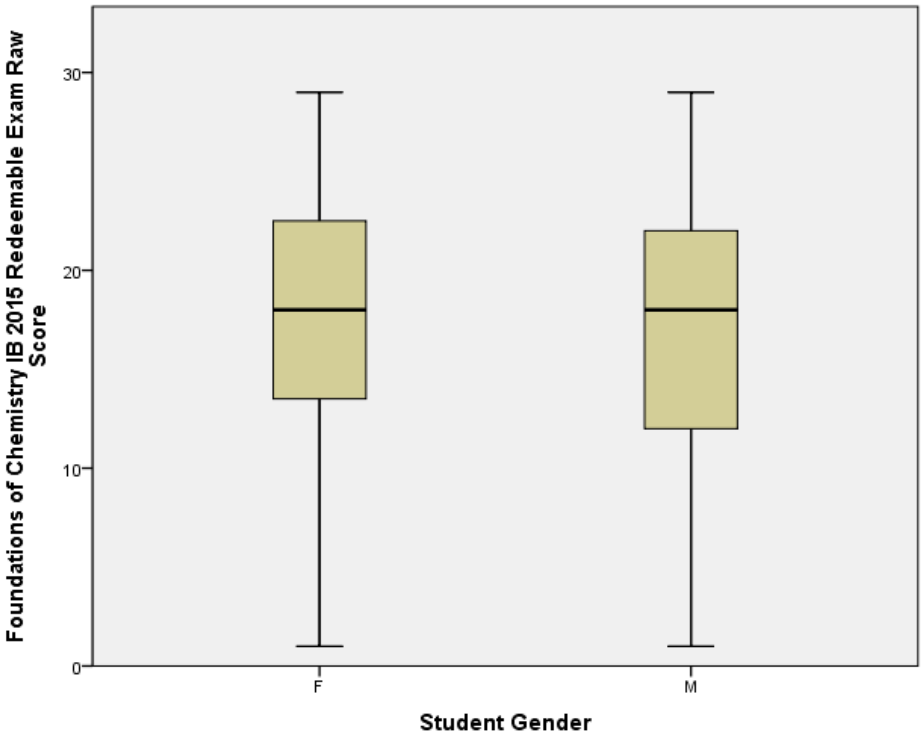

Figure 428: The Boxplot Comparison of Male and Female Raw Score in Foundations of Chemistry IB Redeemable Exam 2015 to Observe Significant Differences

7.12 Gender Bias Items Identified Using of Classical Test Theory

Table 79: Items that were identified to show Significant Differences in the Performance of Male and Female Students through the use of Classical Test Theory within First-Year Chemistry MCQ Assessment Tasks undertaken at The University of Adelaide

	Item	Times Asked	Year	Item	Gender	χ^2	Cohen's *d*
Chemistry IA	**Lec_1_1**	8	2012-2015				
			2012	Lec_1_10	Male	0.003	0.276
			2014	Lec_1_10	Male	0.029	0.201
			2015	Lec_1_10	Male	0.008	0.242
	Lec_1_4	8	2012-2015				
			2012	Lec_1_4	Male	0.004	0.247
			2012	Exam_2_4	Male	0.010	0.220
			2013	Lec_1_4	Male	<0.001	0.363
			2015	Lec_1_4	Male	0.002	0.281
	Lec_1_7	4	2014-2015				
			2015	Lec_1_7	Male	0.010	0.230
			2015	Lec_1_7	Female	0.037	-0.180
	Lec_1_8	6	2012-2015				
			2012	Exam_1_3	Male	<<0.001	0.462
			2013	Exam_1_3	Male	0.024	0.199
			2014	Lec_1_8	Male	0.023	0.210
			2014	Exam_2_8	Male	<<0.001	0.394
			2015	Lec_1_8	Male	0.002	0.286
	Lec_1_9	4	2014-2015				
			2014	Lec_1_9	Male	0.016	0.225
			2014	Exam_2_9	Male	0.001	0.299
			2015	Lec_1_9	Male	0.001	0.286
	Lec_1_10	4	2012-2015				
			2014	Lec_1_1	Male	0.036	0.192
			2015	Lec_1_1	Male	0.007	0.240
	Lec_2_2	8	2012-2015				
			2012	Exam_2_17	Male	0.010	0.220
			2013	Lec_2_2	Male	0.001	0.341
			2014	Exam_2_17	Male	0.012	0.223
			2015	Lec_2_2	Male	<0.001	0.354
	Lec_2_3	4	2012-2015				
			2014	Lec_2_3	Male	0.022	0.221
			2015	Lec_2_3	Male	0.014	0.235
	Lec_2_4	8	2012-2015				
			2012	Lec_2_4	Male	0.011	0.244
			2014	Lec_2_4	Male	0.001	0.305
			2015	Lec_2_4	Male	0.003	0.292

	Lec_2_12	4	2012-2015				
			2014	Lec_2_12	Male	0.046	0.191
			2015	Lec_2_12	Male	0.031	0.205
	Exam_1_1	4	2012-2015				
			2013	Exam_1_1	Female	0.006	-0.240
			2014	Exam_1_5	Female	0.030	-0.193
	Exam_2_19	8	2012-2015				
			2012	Exam_2_19	Male	0.023	0.199
			2013	Exam_2_19	Female	0.003	-0.270
			2015	Exam_2_19	Female	0.006	-0.234
Chemistry IB	**Lec_1_3**	8	2012-2015				
			2012	Lec_1_3	Male	0.014	0.251
			2015	Lec_1_3	Male	0.019	0.230
	Lec_1_8	8	2012-2015				
			2012	Lec_1_8	Male	<<0.001	0.440
			2013	Lec_1_8	Male	0.002	0.320
	Lec_1_11	8	2012-2015				
			2013	Lec_1_11	Male	0.002	0.326
			2013	Exam_2_11	Male	0.029	0.206
			2014	Lec_1_11	Male	0.014	0.233
			2015	Exam_2_11	Male	0.008	0.242
	Lec_1_12	8	2012-2015				
			2012	Lec_1_12	Male	0.003	0.312
			2012	Exam_2_12	Male	0.007	0.261
			2013	Lec_1_12	Male	0.009	0.270
			2013	Exam_2_12	Male	0.001	0.315
			2014	Exam_2_12	Male	0.002	0.286
	Lec_2_4	4	2012-2015				
			2014	Lec_2_4	Male	0.003	0.302
			2015	Lec_2_4	Male	0.009	0.268
	Lec_2_15	8	2012-2015				
			2013	Lec_2_15	Male	0.028	0.245
			2013	Lec_2_29	Female	0.003	-0.287
	Exam_1_9	4	2012-2015				
			2012	Exam_1_9	Male	0.047	0.191
			2015	Exam_1_9	Male	0.015	0.221
	Exam_2_4	8	2012-2015				
			2013	Exam_2_4	Male	0.050	0.185
			2014	Exam_2_4	Male	0.040	0.188
			2015	Exam_2_4	Male	0.009	0.268
	Exam_2_25	8	2012-2015				
			2012	Exam_2_25	Female	0.006	-0.265
			2015	Exam_2_25	Female	0.001	-0.292

Foundations of Chemistry IA	**Lec_2_10**	4	2012-2015				
			2014	Lec_2_10	Female	0.020	-0.315
			2015	Lec_2_10	Male	0.022	0.303
	Exam_1_2	4	2012-2015				
			2012	Exam_1_2	Male	0.036	0.213
			2013	Exam_1_2	Male	0.031	0.216
	Exam_2_19	4	2012-2015				
			2014	Exam_2_19	Female	0.015	-0.270
			2015	Exam_2_19	Female	0.039	-0.216
Foundations of Chemistry IB	**Lec_1_14**	4	2012-2015				
			2013	Lec_1_14	Female	0.047	-0.252
			2015	Lec_1_14	Male	0.034	0.281
	Exam_1_8	1	2012				
			2012	Exam_1_8	Male	0.006	0.349
	Exam_2_11	4	2012-2015				
			2013	Exam_2_11	Female	0.019	-0.249
			2014	Exam_2_11	Female	0.024	-0.269

7.13 Comparison of Male and Female Student Ability in Chemistry MCQ Assessments

Table 80: The Student Ability Comparison of Male and Female Students in Chemistry IA MCQ Assessments to Determine if a Difference Should be Expected in how each Gender Answers Individual Items (All Significance Favours Male Students)

Chemistry IA	**2012**		**2013**		**2014**		**2015**	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	467	0.008	446	0.184	471	0.069	502	0
Lecture Test 2	444	0.093	418	0.012	434	0.01	449	0
Exam	506	0.002	503	0.983	505	0.355	544	0.707
Redeemable Exam	485	0.165	486	0.244	496	0.001	523	0.382

Table 81: The Student Ability Comparison of Male and Female Students in Chemistry I**B** MCQ Assessments to Determine if a Difference Should be Expected in how each Gender Answers Individual Items (All Significance Favours Male Students)

Chemistry IB	**2012**		**2013**		**2014**		**2015**	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	380	0.024	376	0.007	421	0.019	426	0.001
Lecture Test 2	362	0.182	346	0.136	392	0.101	389	<<0.001
Exam	431	0.683	448	0.71	484	0.127	484	0.303
Redeemable Exam	419	0.277	432	0.513	454	0.295	469	0.193

Table 82: The Student Ability Comparison of Male and Female Students in Foundations of Chemistry IA MCQ Assessments to Determine if a Difference Should be Expected in how each Gender Answers Individual Items (All Significance Favours Male Students)

Foundations of Chemistry IA	**2012**		**2013**		**2014**		**2015**	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	257	0.389	307	0.319	250	0.892	292	0.098
Lecture Test 2	265	0.028	253	0.837	221	0.358	234	0.04
Exam	304	0.439	363	0.133	325	0.69	365	0.756
Redeemable Exam	256	0.632	334	0.344	299	0.691	329	0.116

Table 83: The Student Ability Comparison of Male and Female Students in Foundations of Chemistry IB MCQ Assessments to Determine if a Difference Should be Expected in how each Gender Answers Individual Items (All Significance Favours Male Students)

Foundations of Chemistry IB	**2012**		**2013**		**2014**		**2015**	
	d.f.	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**	**d.f.**	***p*-value**
Lecture Test 1	234	0.56	247	0.94	214	0.896	229	0.05
Lecture Test 2	187	0.632	216	0.887	182	0.707	196	0.409
Exam	264	0.103	303	0.292	274	0.465	298	0.026
Redeemable Exam	248	0.985	286	0.685	257	0.994	275	0.792

7.14 Boxplot Comparison of Male and Female Student Ability in Chemistry MCQ Assessments

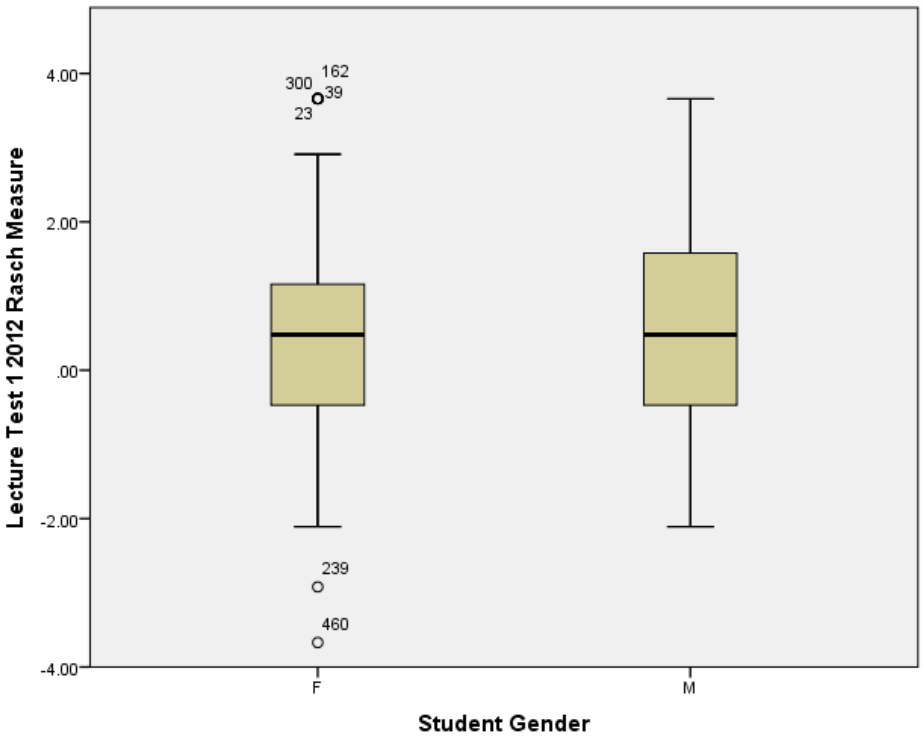

Figure 429: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Lecture Test 1 2012 to Observe Significant Differences

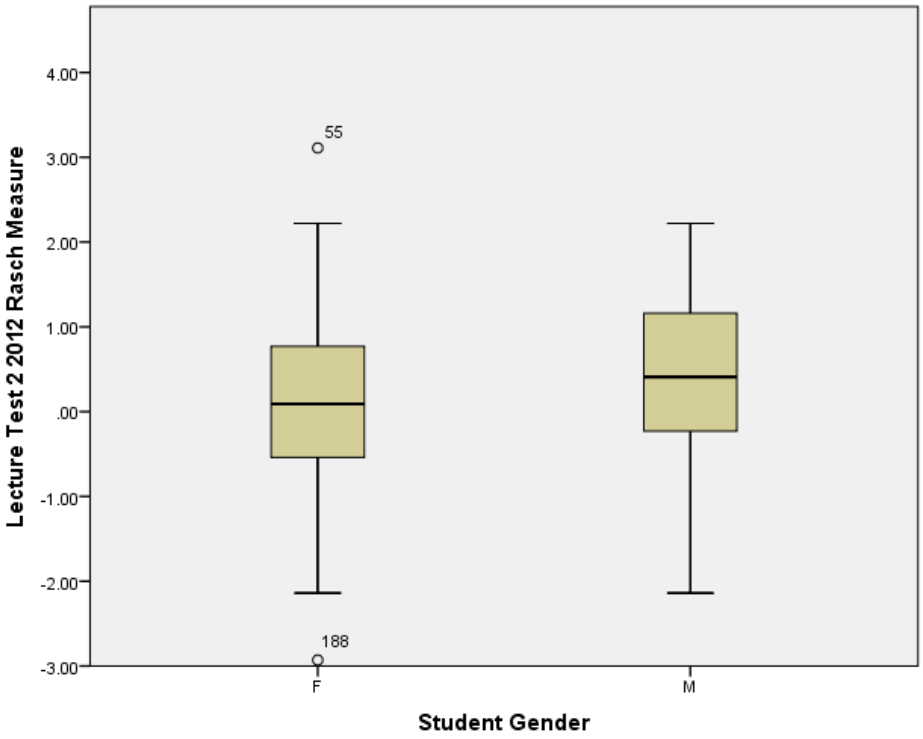

Figure 430: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Lecture Test 2 2012 to Observe Significant Differences

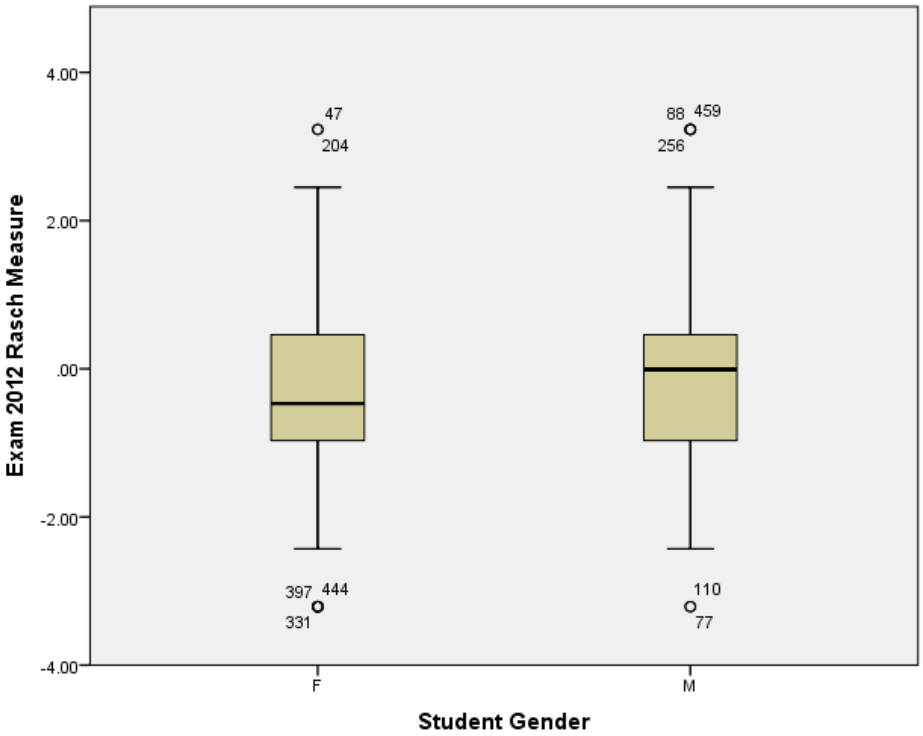

Figure 431: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Exam 2012 to Observe Significant Differences

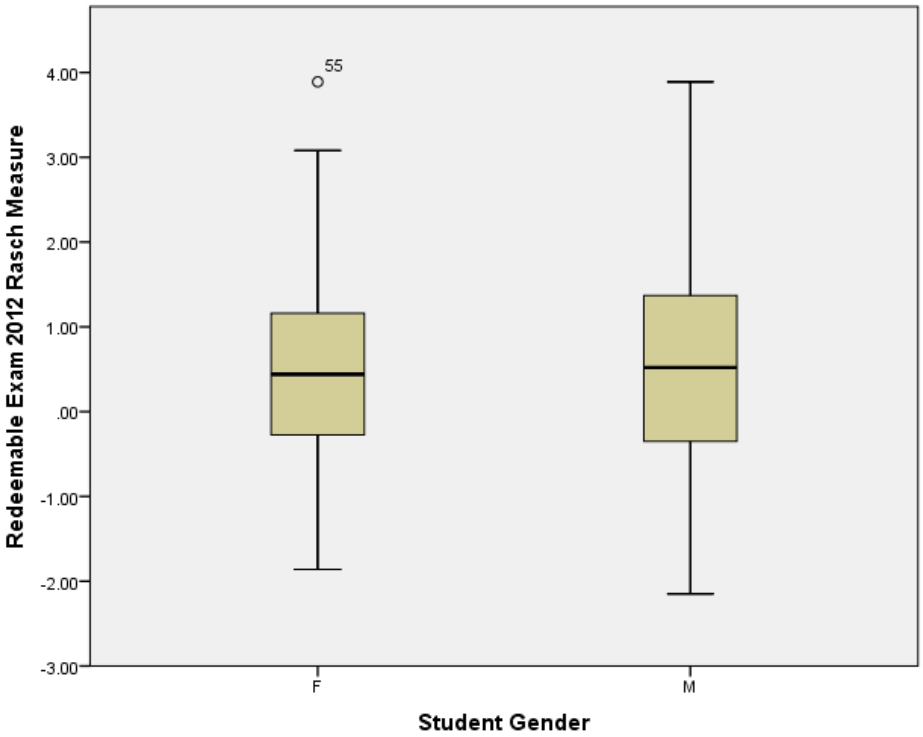

Figure 432: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Redeemable Exam 2012 to Observe Significant Differences

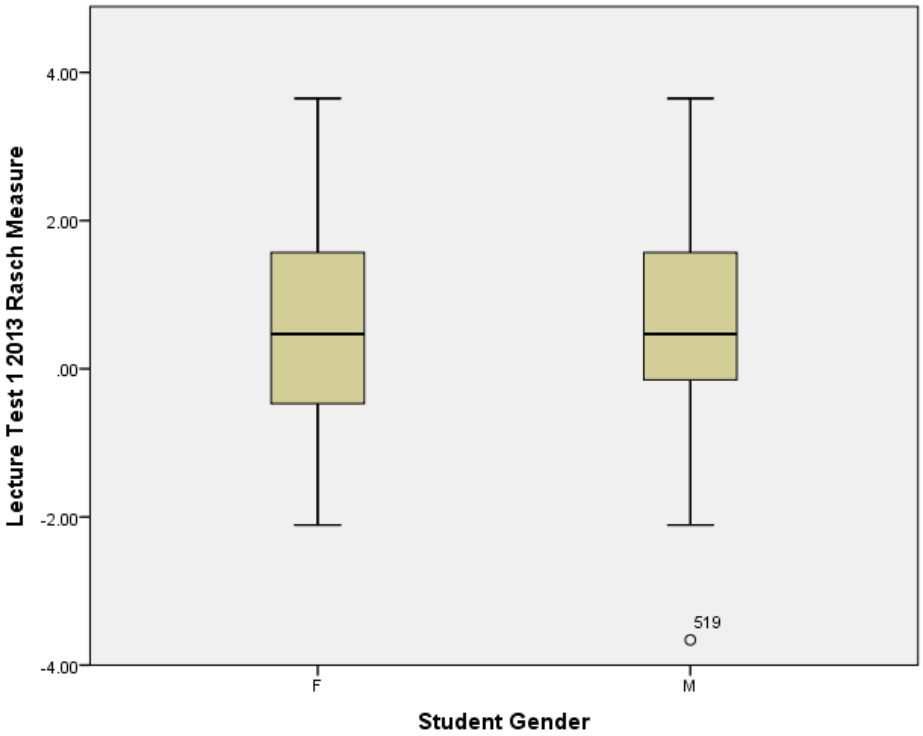

Figure 433: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Lecture Test 1 2013 to Observe Significant Differences

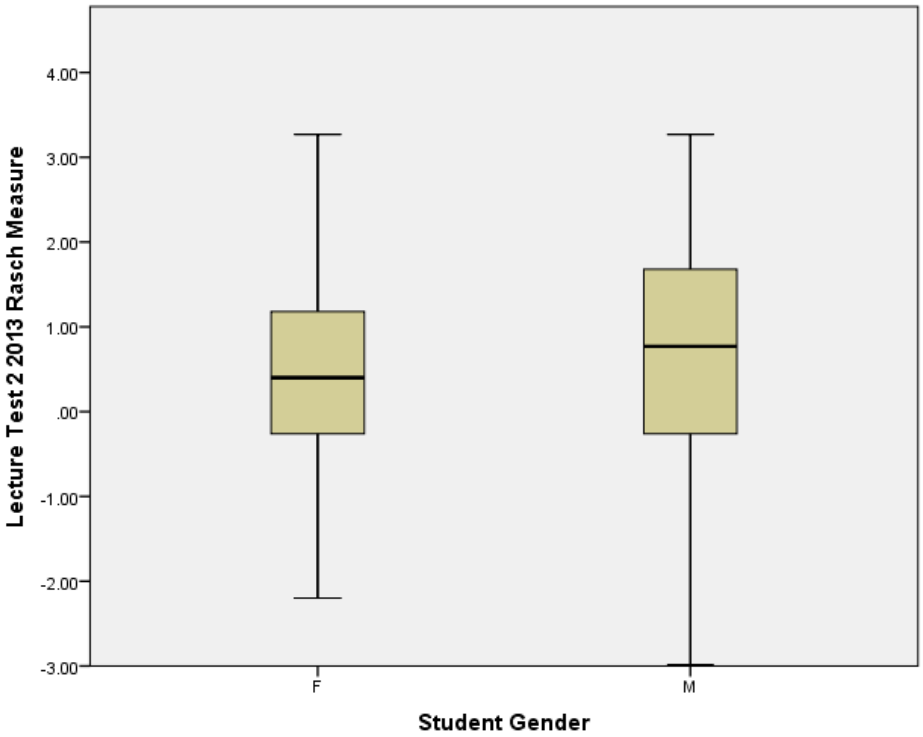

Figure 434: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Lecture Test 2 2013 to Observe Significant Differences

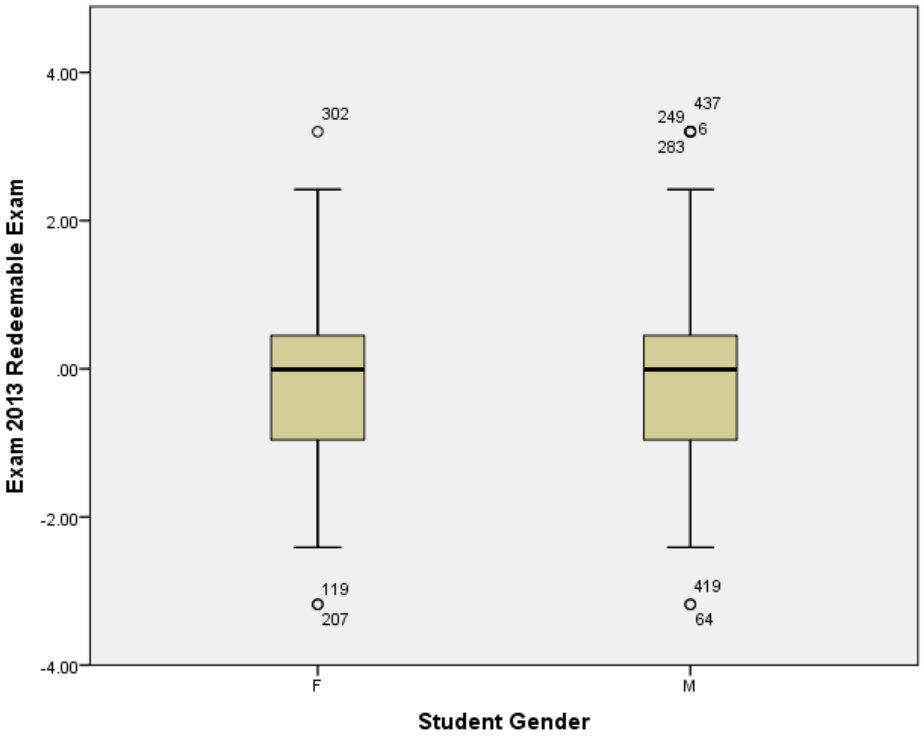

Figure 435: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Exam 2013 to Observe Significant Differences

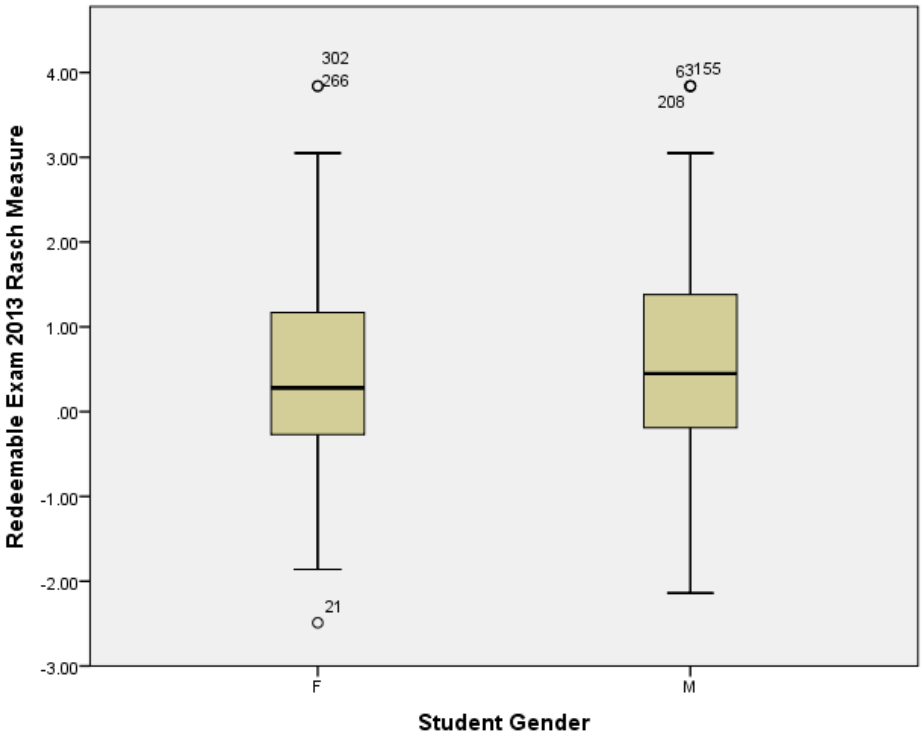

Figure 436: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Redeemable Exam 2013 to Observe Significant Differences

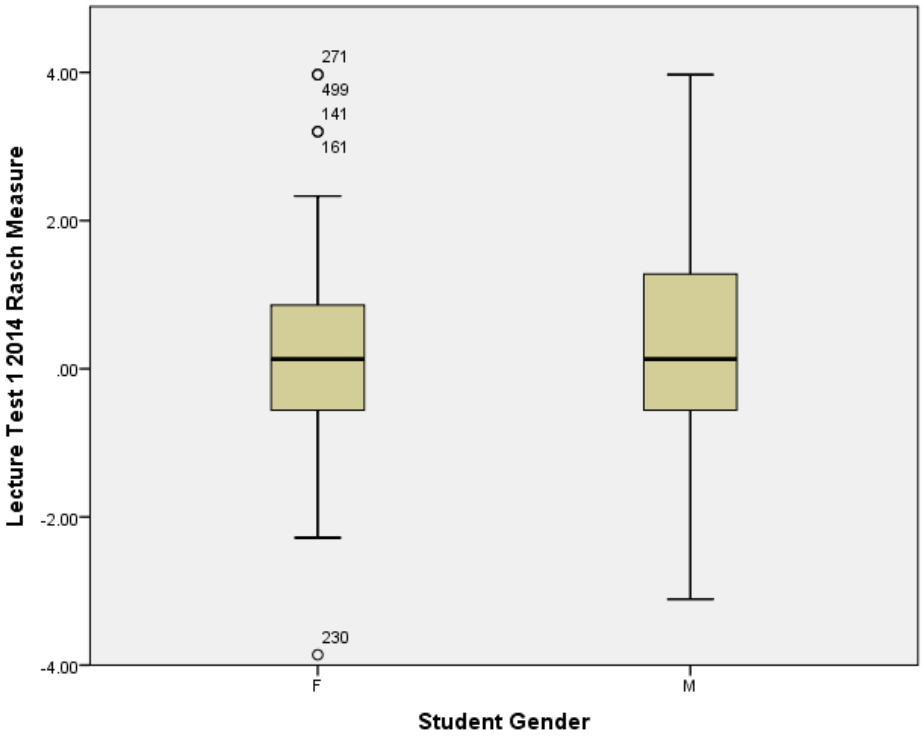

Figure 437: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Lecture Test 1 2014 to Observe Significant Differences

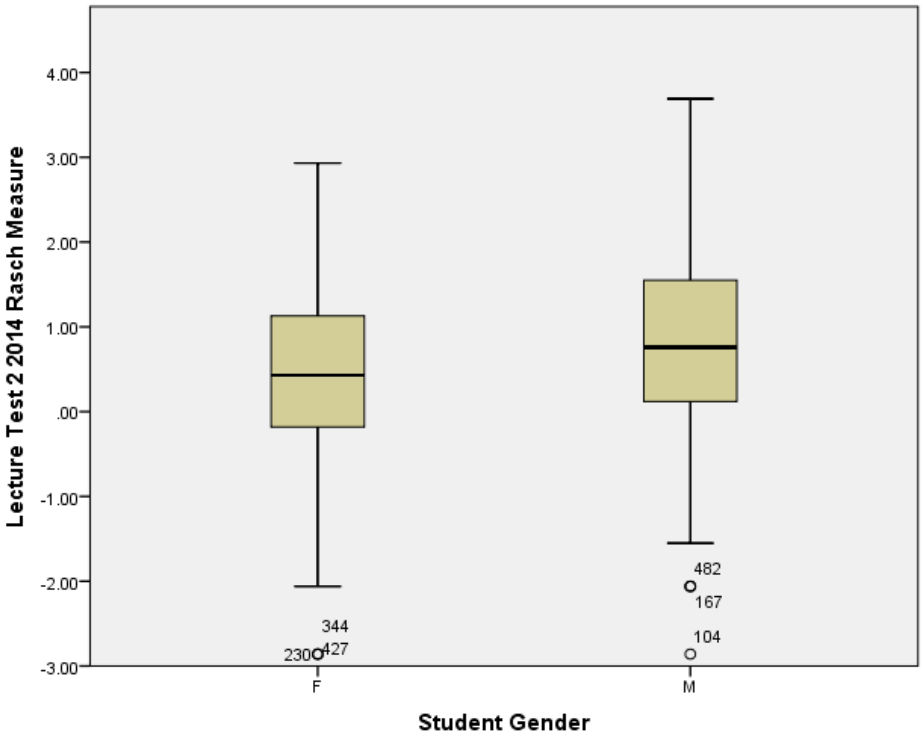

Figure 438: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Lecture Test 2 2014 to Observe Significant Differences

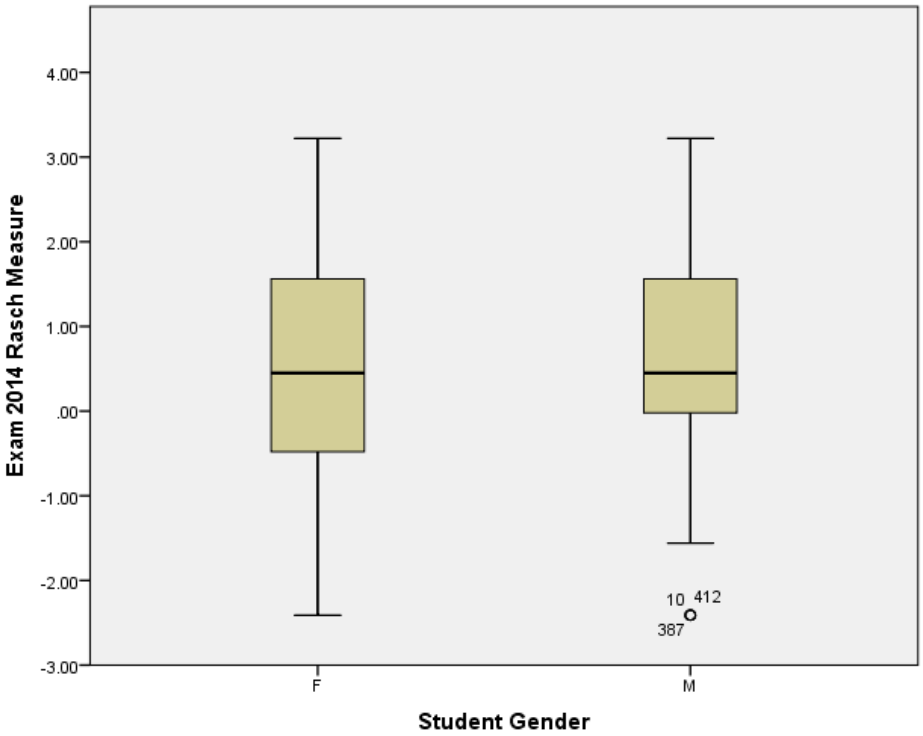

Figure 439: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Exam 2014 to Observe Significant Differences

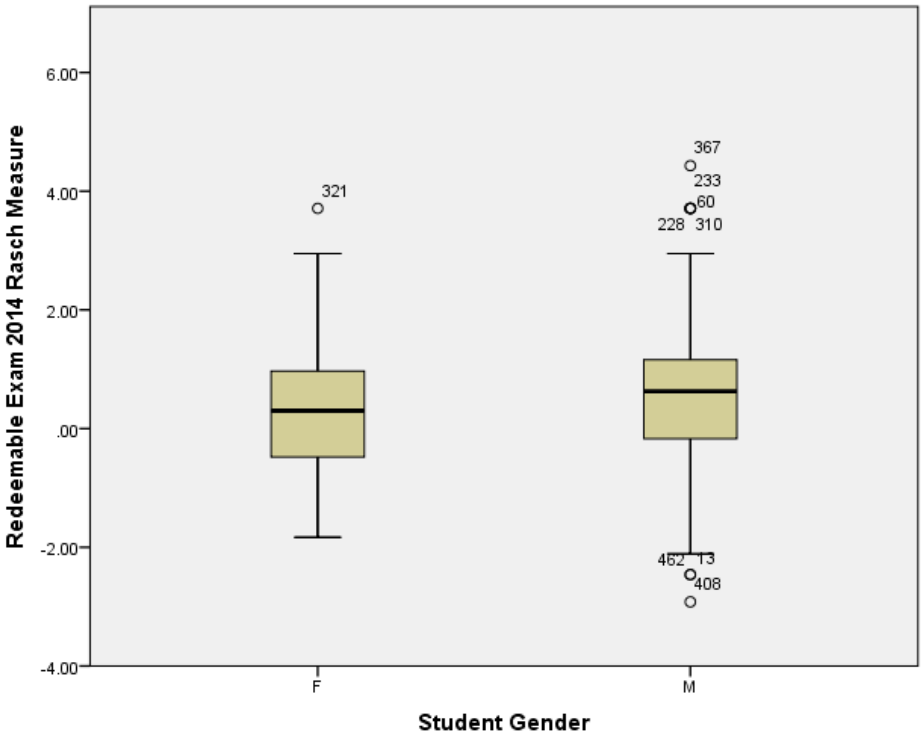

Figure 440: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Redeemable Exam 2014 to Observe Significant Differences

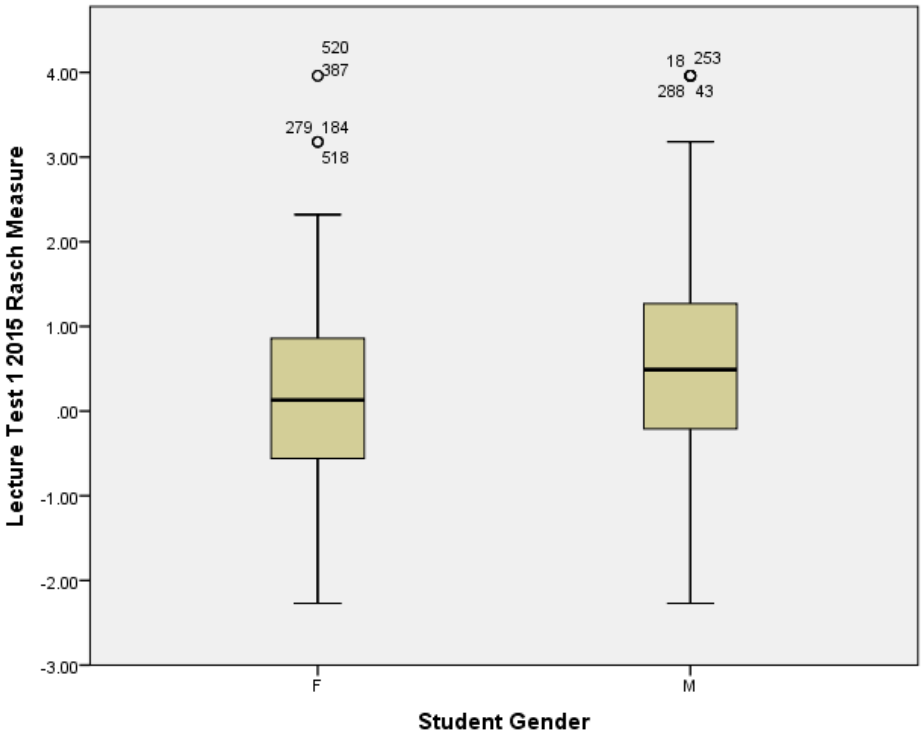

Figure 441: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Lecture Test 1 2015 to Observe Significant Differences

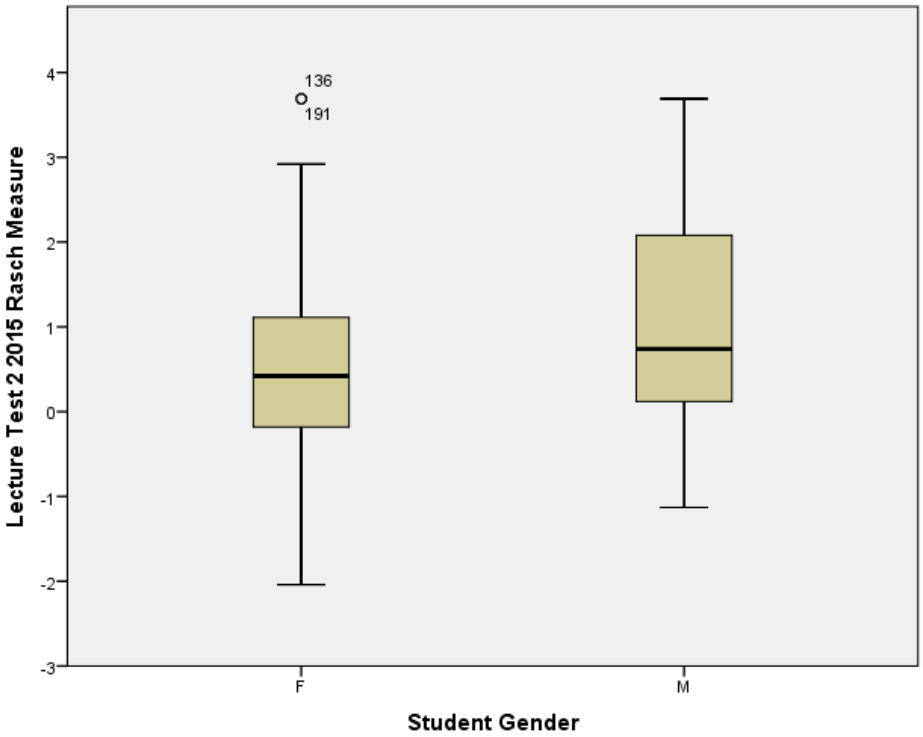

Figure 442: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Lecture Test 2 2015 to Observe Significant Differences

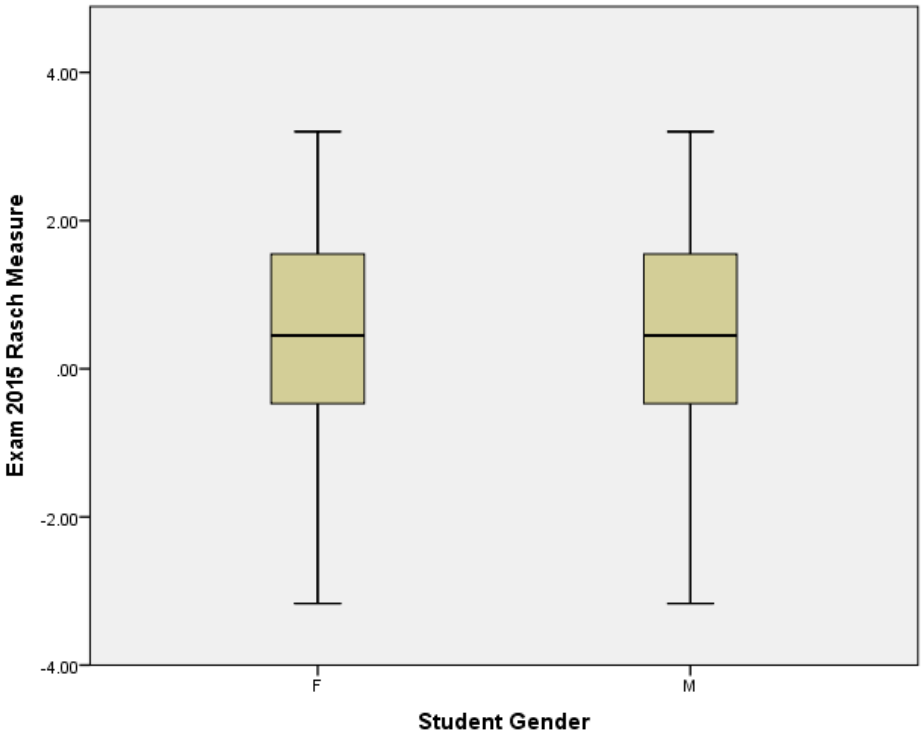

Figure 443: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Exam 2015 to Observe Significant Differences

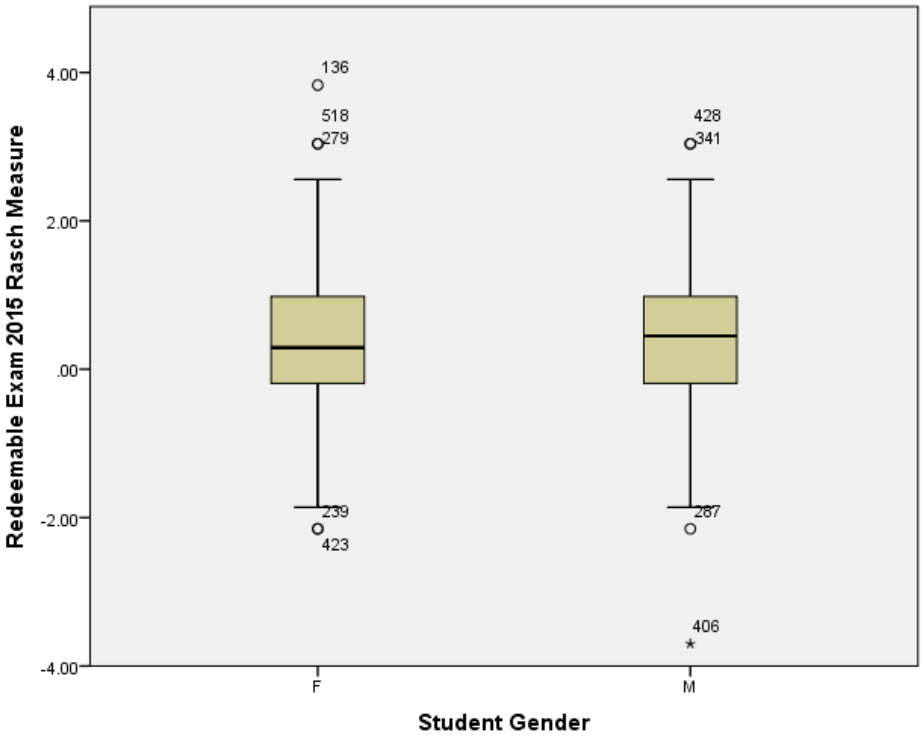

Figure 444: The Boxplot Comparison of Male and Female Student Ability in Chemistry IA Redeemable Exam 2015 to Observe Significant Differences

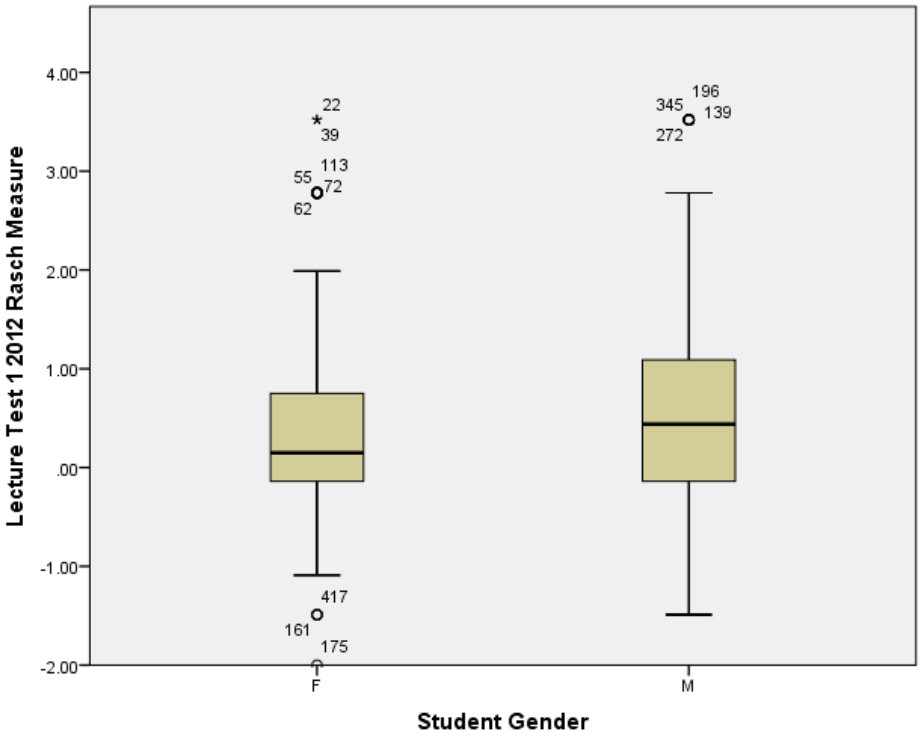

Figure 445: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Lecture Test 1 2012 to Observe Significant Differences

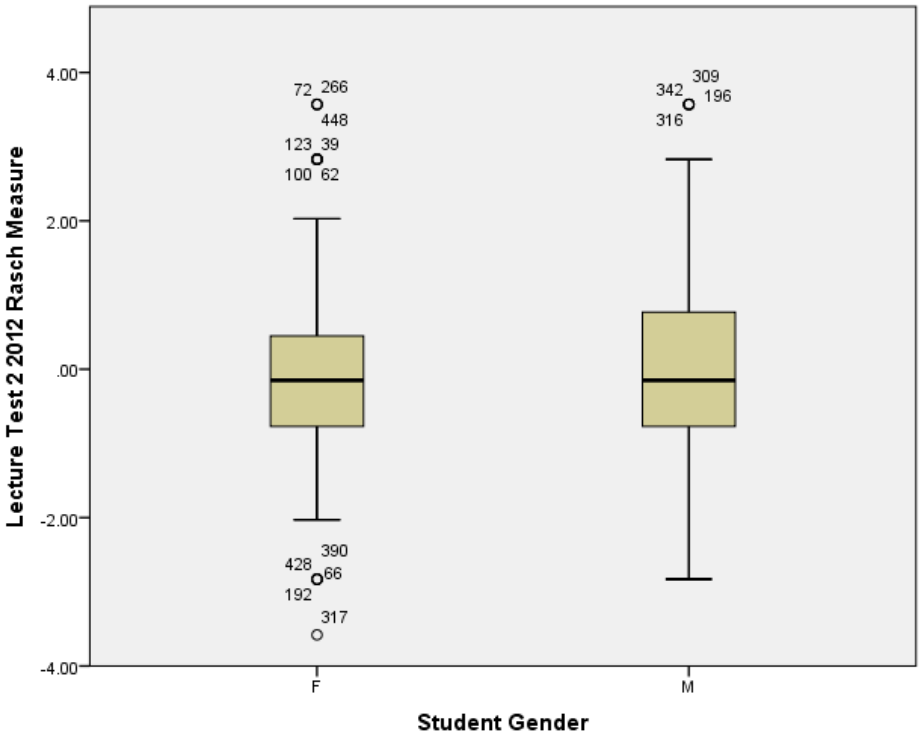

Figure 446: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Lecture Test 2 2012 to Observe Significant Differences

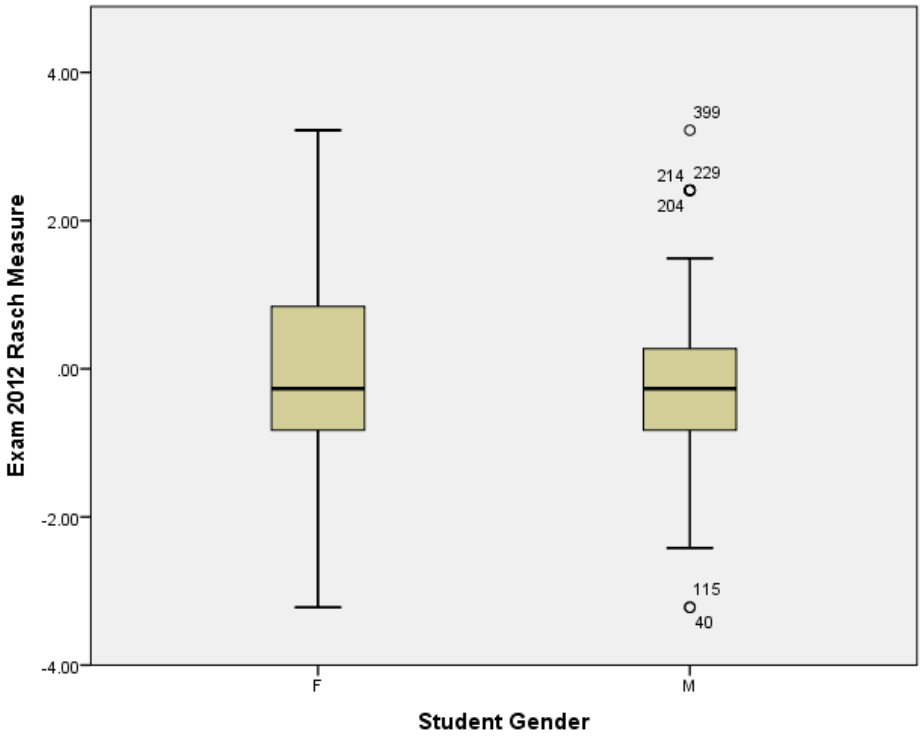

Figure 447: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Exam 2012 to Observe Significant Differences

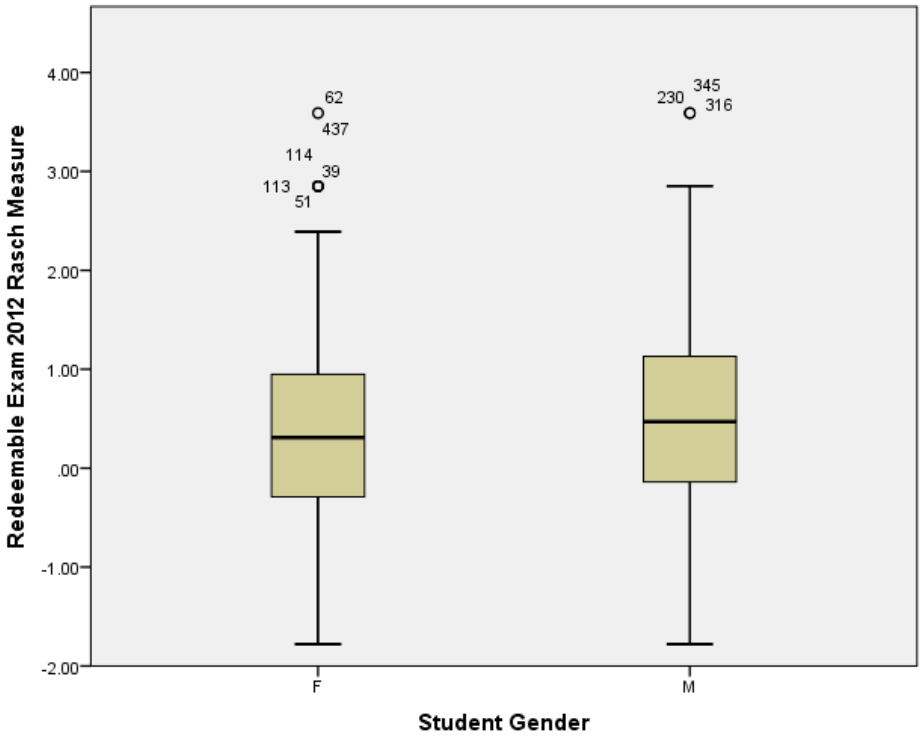

Figure 448: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Redeemable Exam 2012 to Observe Significant Differences

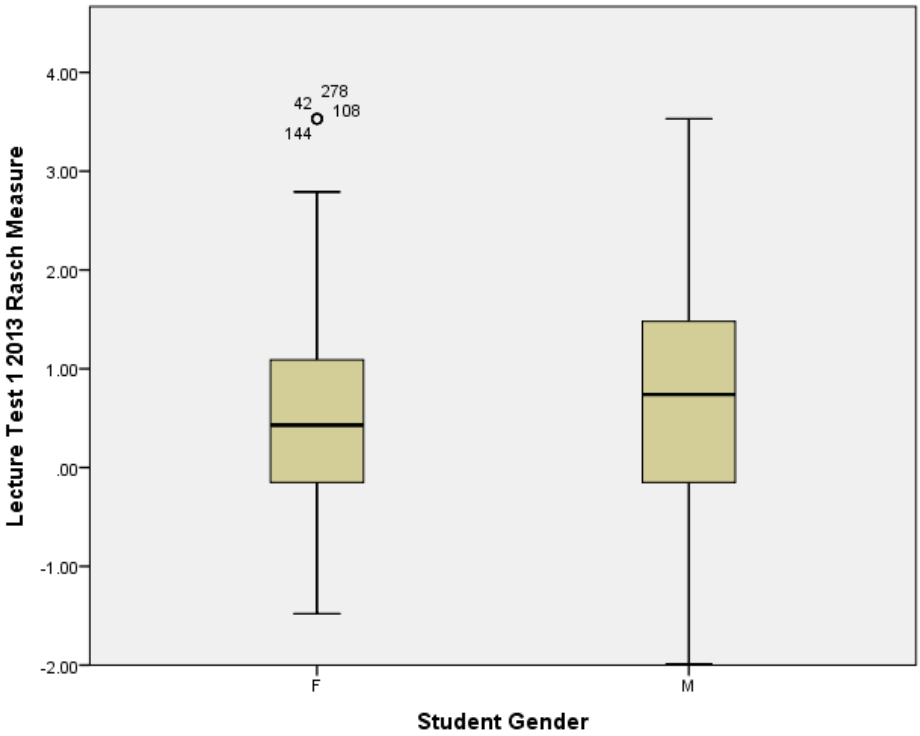

Figure 449: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Lecture Test 1 2013 to Observe Significant Differences

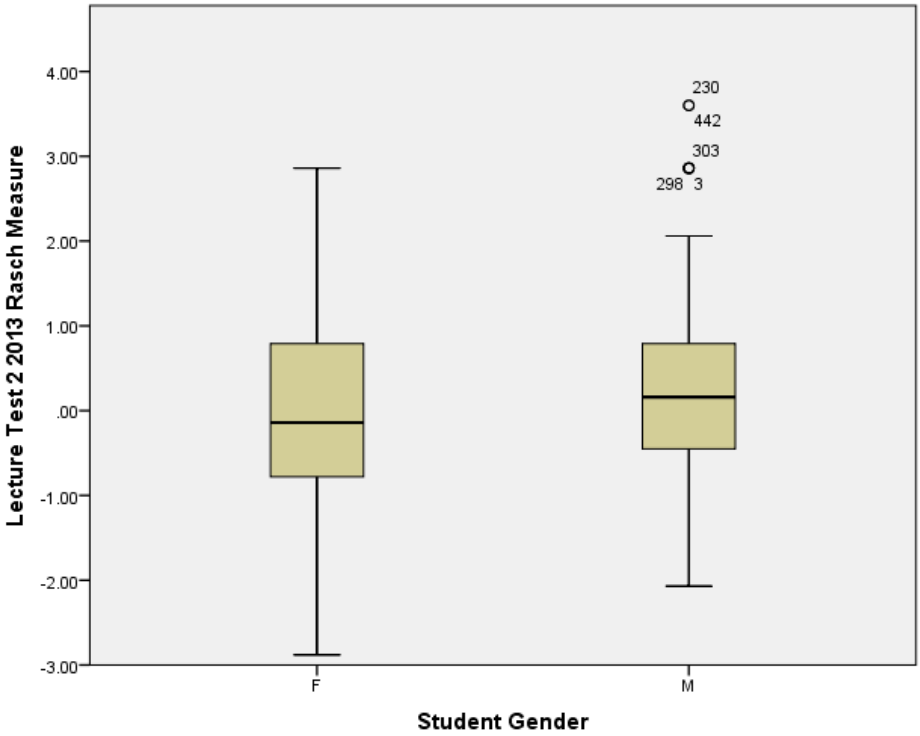

Figure 450: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Lecture Test 2 2013 to Observe Significant Differences

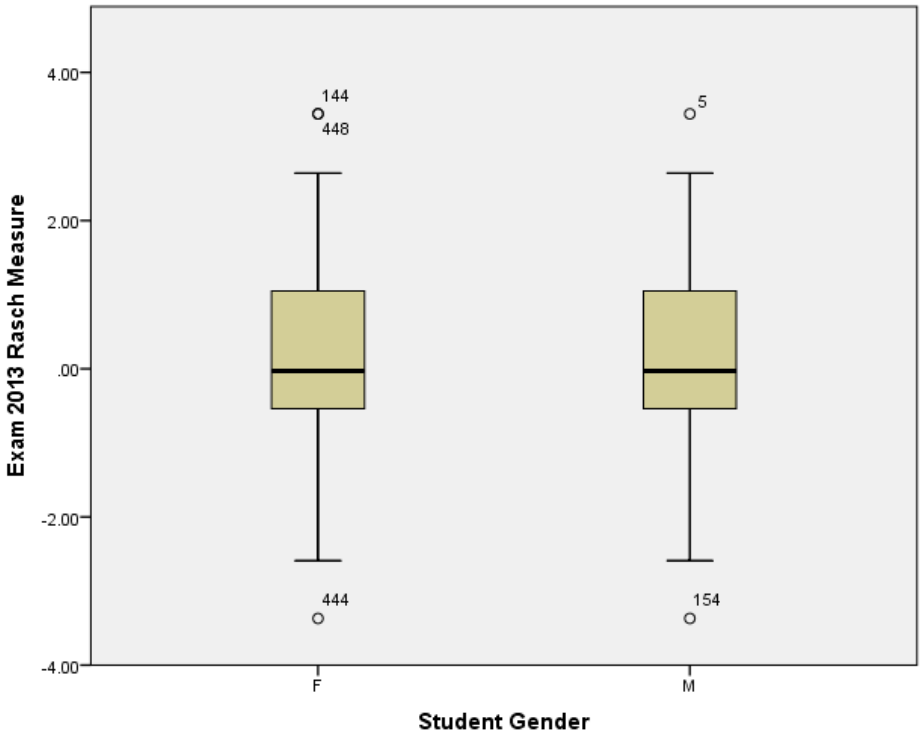

Figure 451: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Exam 2013 to Observe Significant Differences

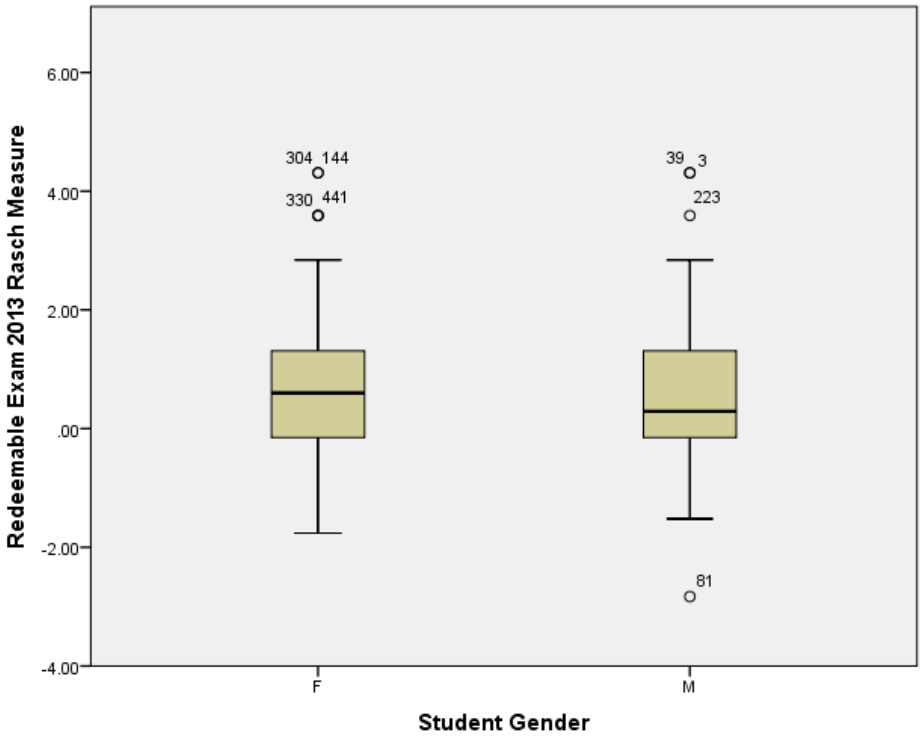

Figure 452: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Redeemable Exam 2013 to Observe Significant Differences

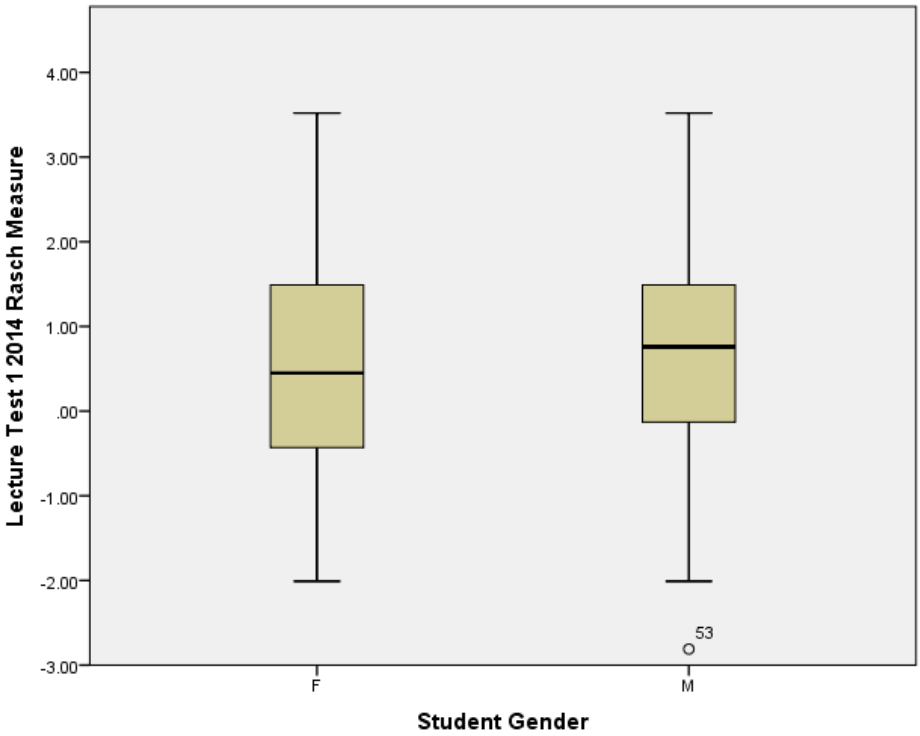

Figure 453: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Lecture Test 1 2014 to Observe Significant Differences

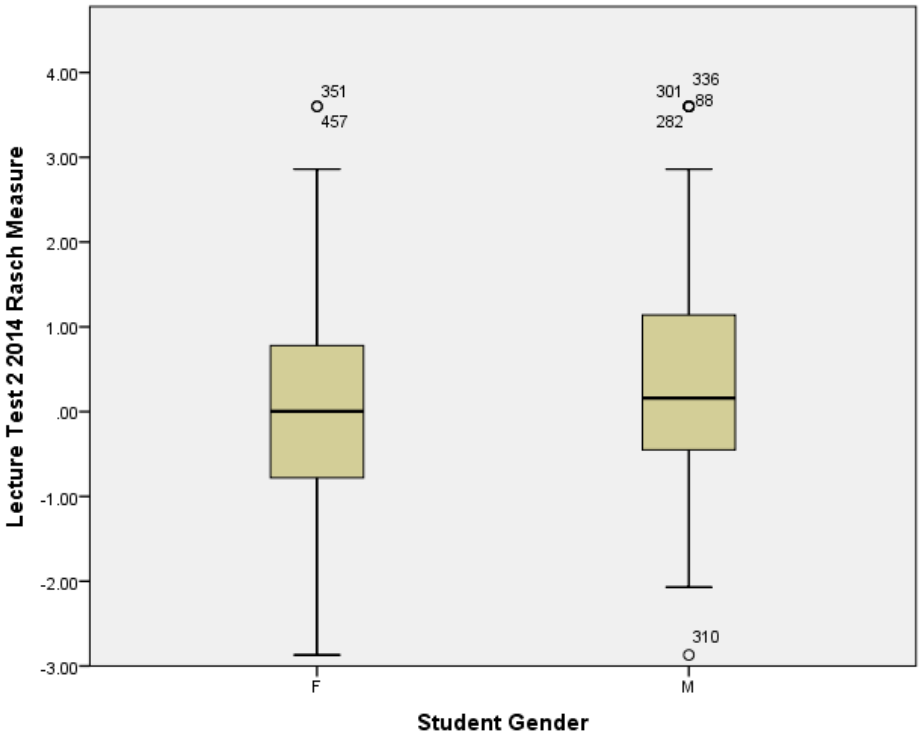

Figure 454: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Lecture Test 2 2014 to Observe Significant Differences

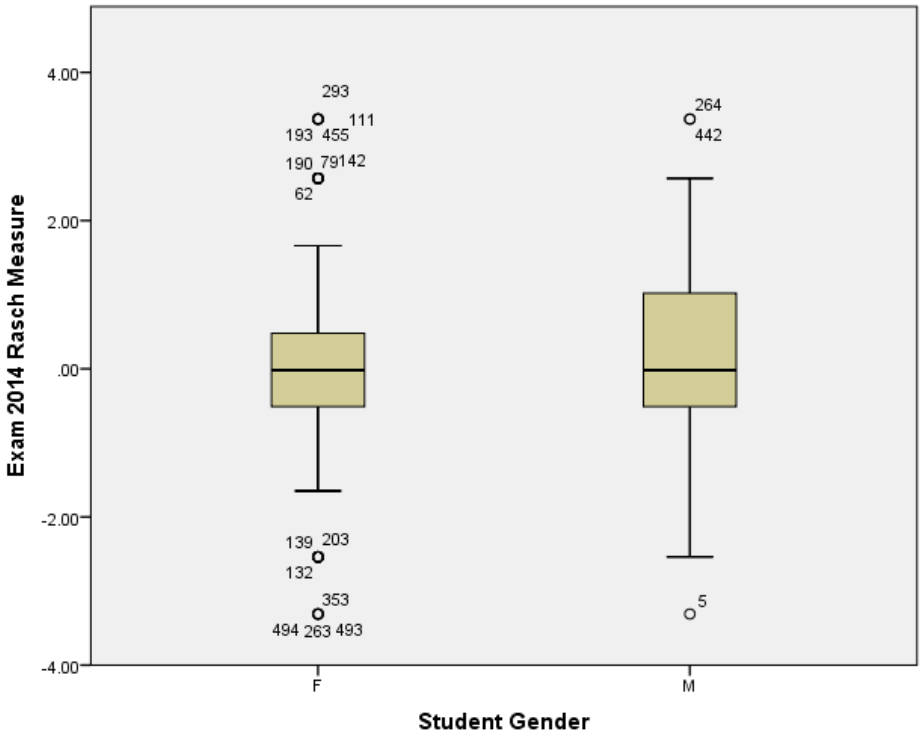

Figure 455: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Exam 2014 to Observe Significant Differences

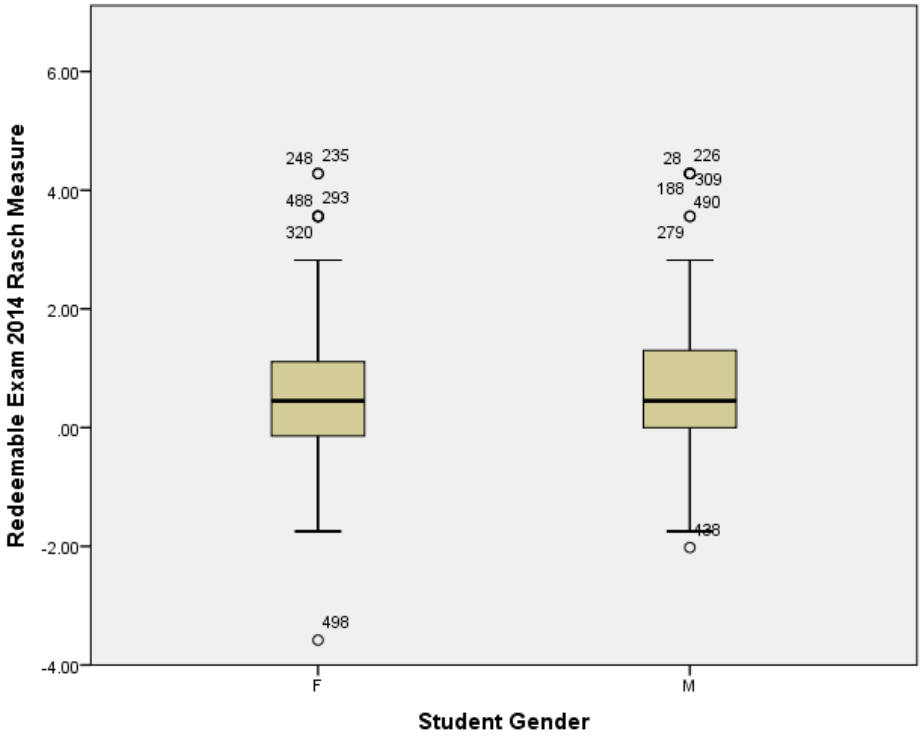

Figure 456: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Redeemable Exam 2014 to Observe Significant Differences

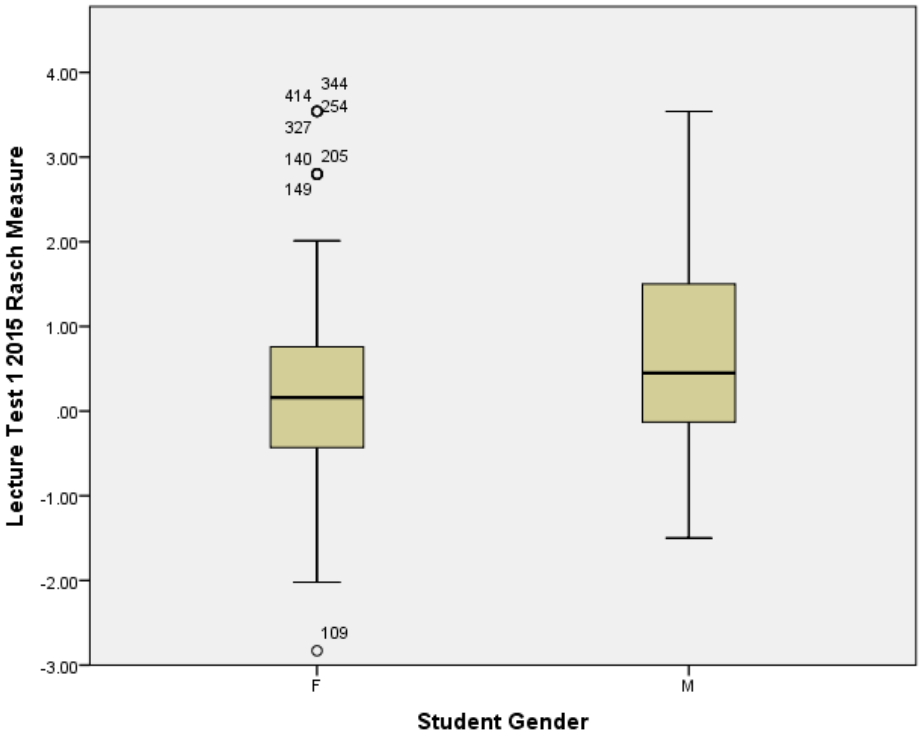

Figure 457: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Lecture Test 1 2015 to Observe Significant Differences

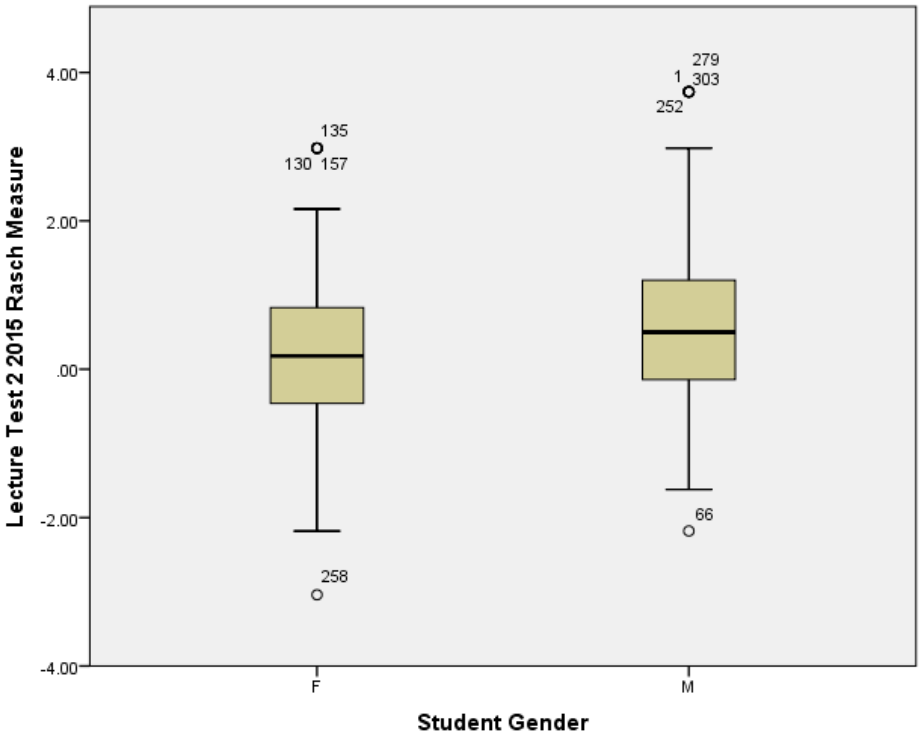

Figure 458: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Lecture Test 2 2015 to Observe Significant Differences

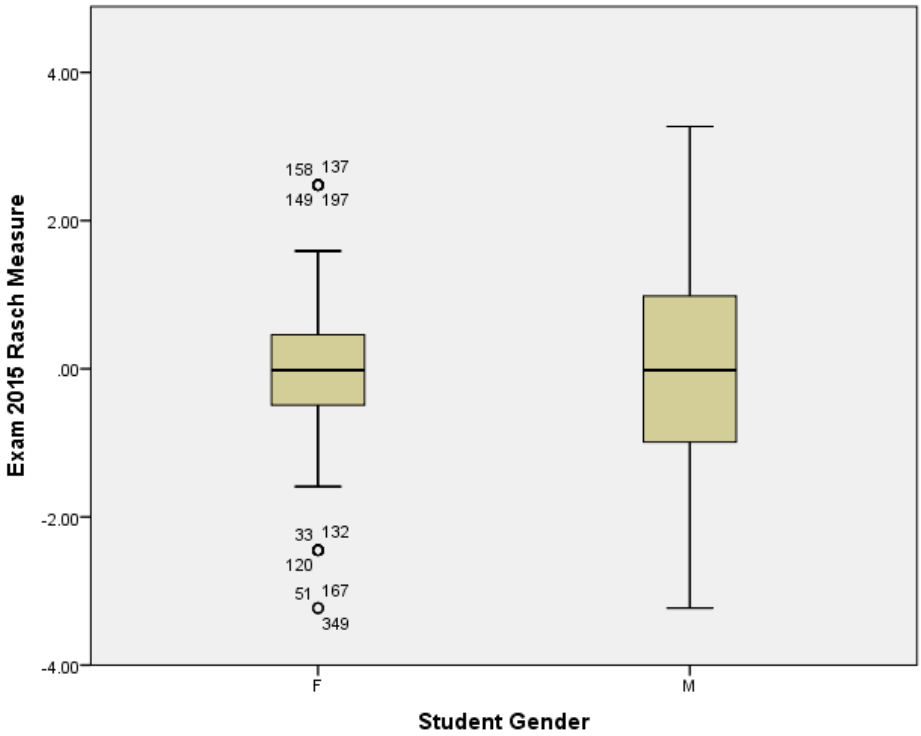

Figure 459: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Exam 2015 to Observe Significant Differences

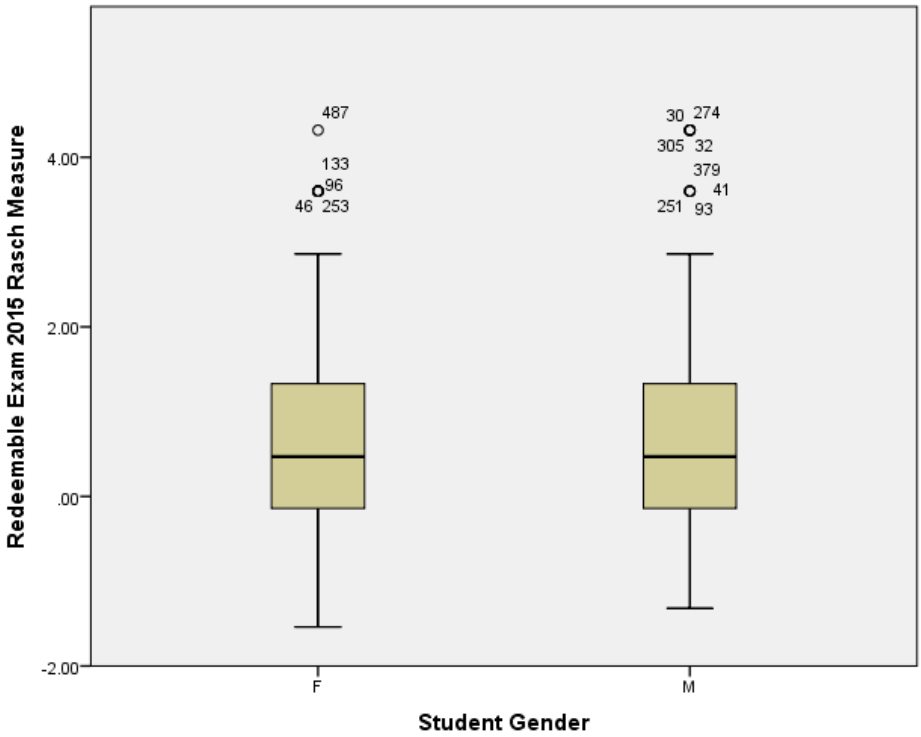

Figure 460: The Boxplot Comparison of Male and Female Student Ability in Chemistry IB Redeemable Exam 2015 to Observe Significant Differences

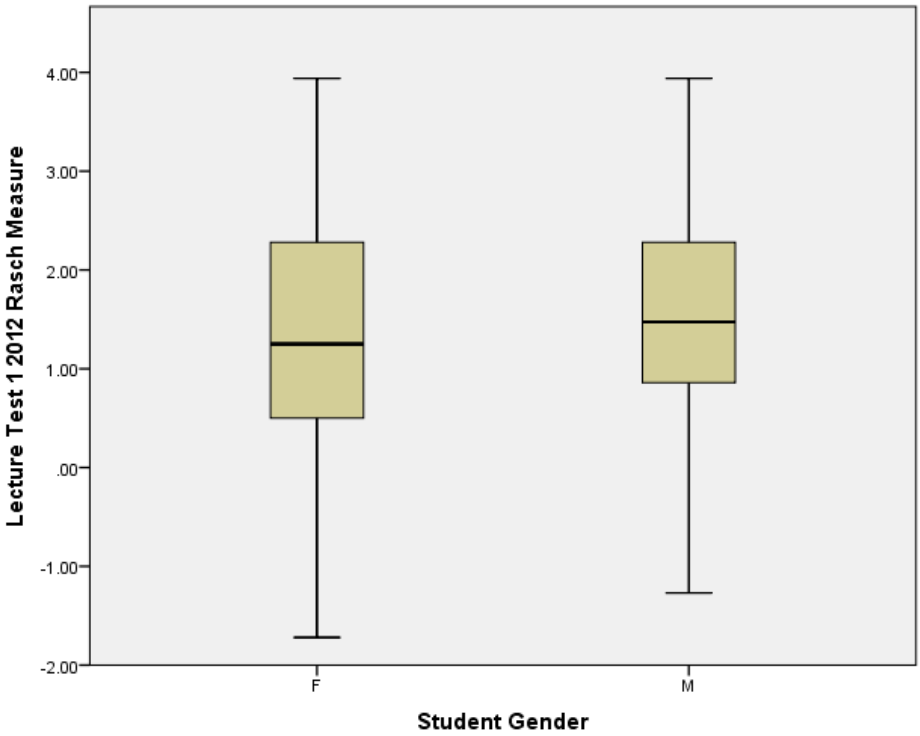

Figure 461: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Lecture Test 1 2012 to Observe Significant Differences

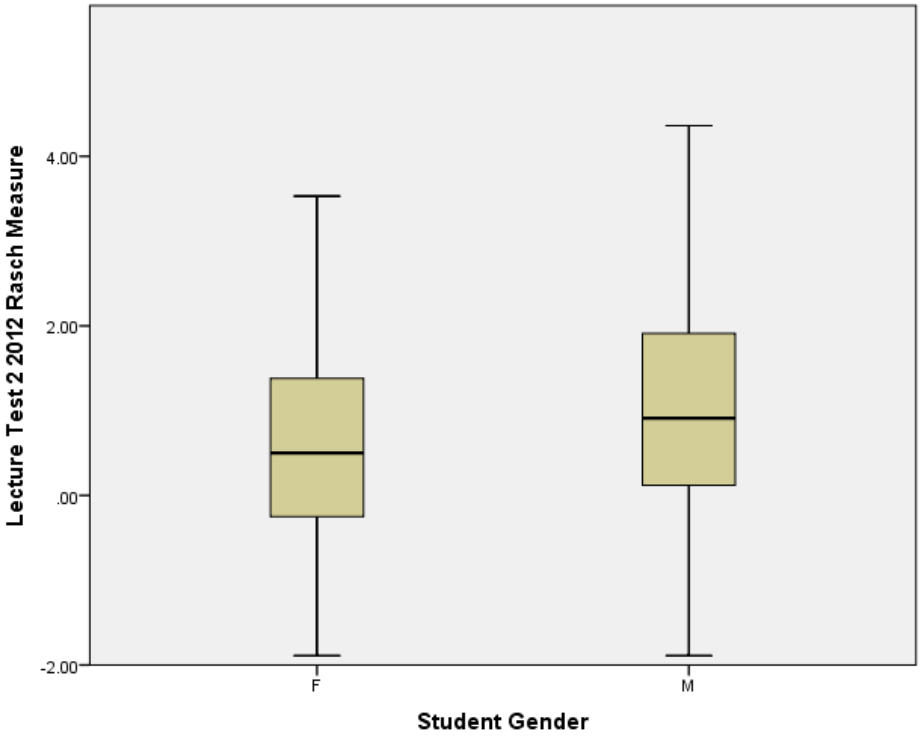

Figure 462: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Lecture Test 2 2012 to Observe Significant Differences

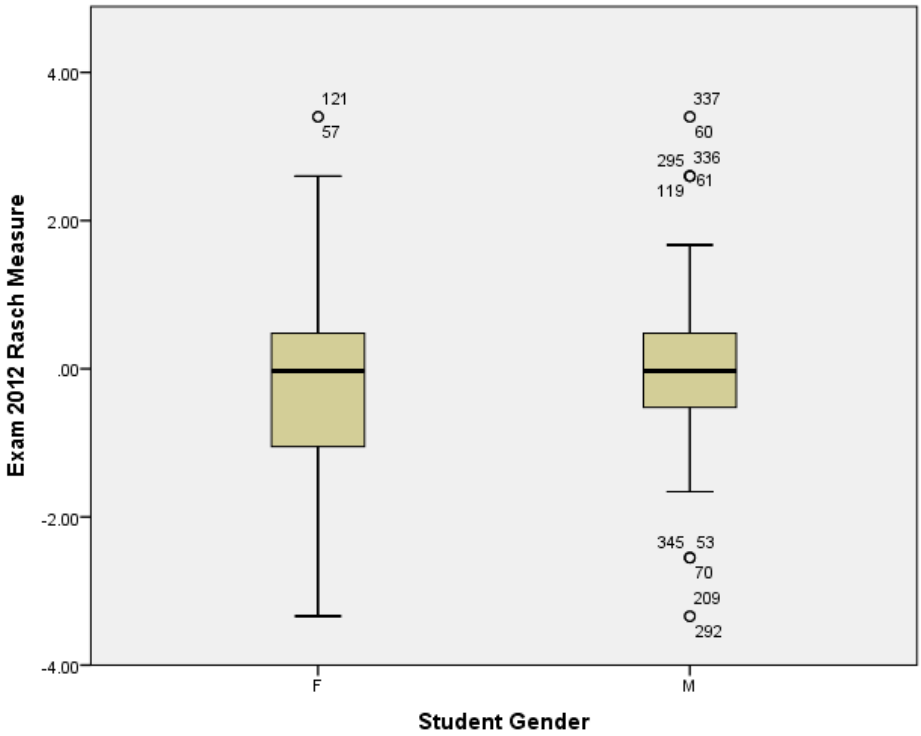

Figure 463: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Exam 2012 to Observe Significant Differences

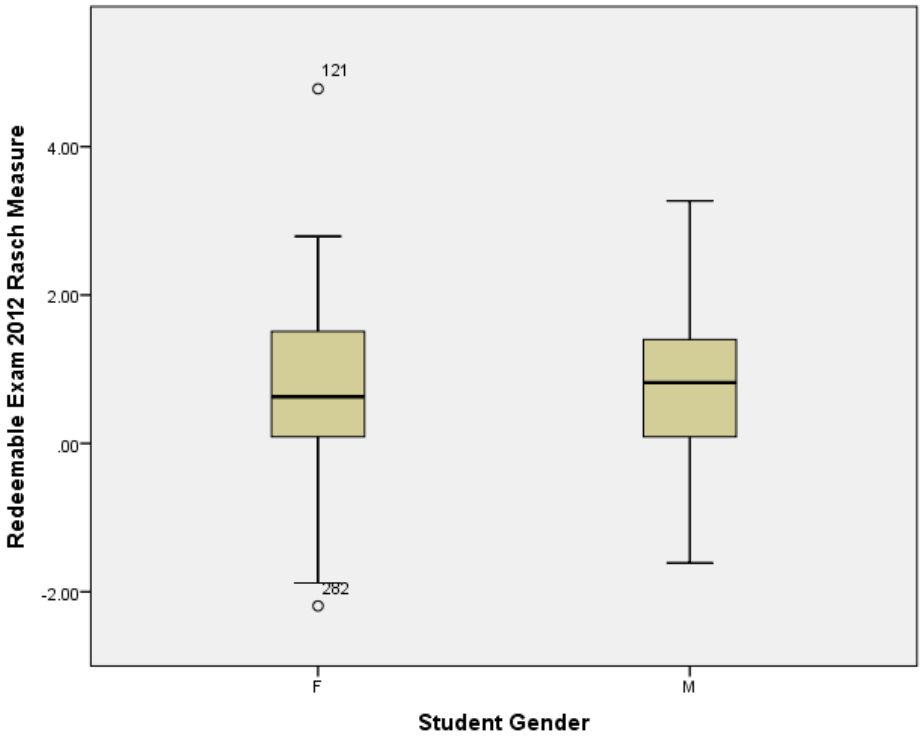

Figure 464: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Redeemable Exam 2012 to Observe Significant Differences

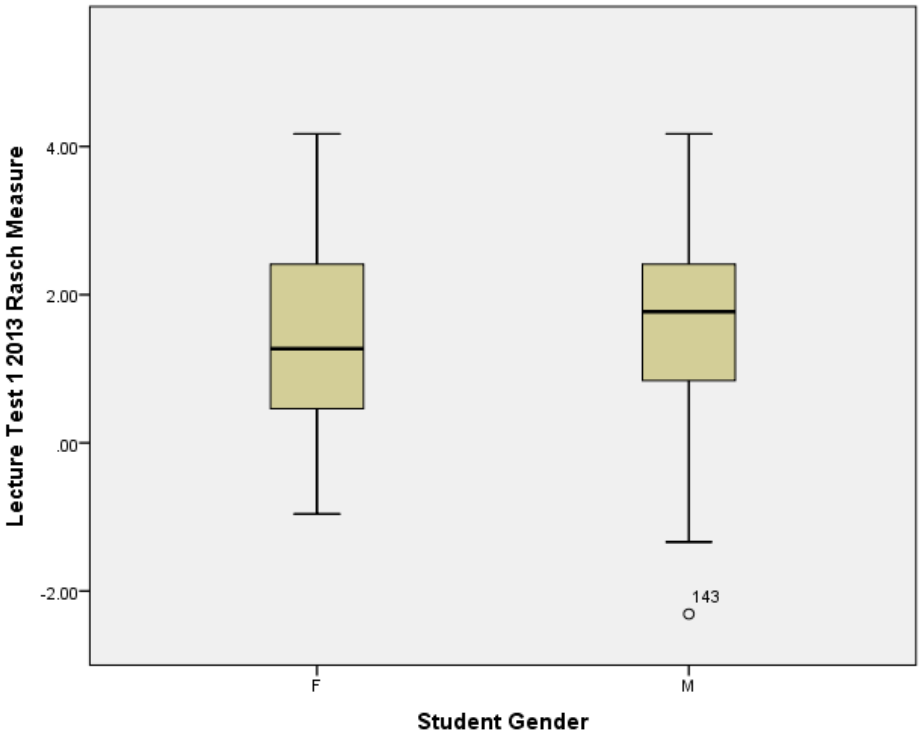

Figure 465: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Lecture Test 1 2013 to Observe Significant Differences

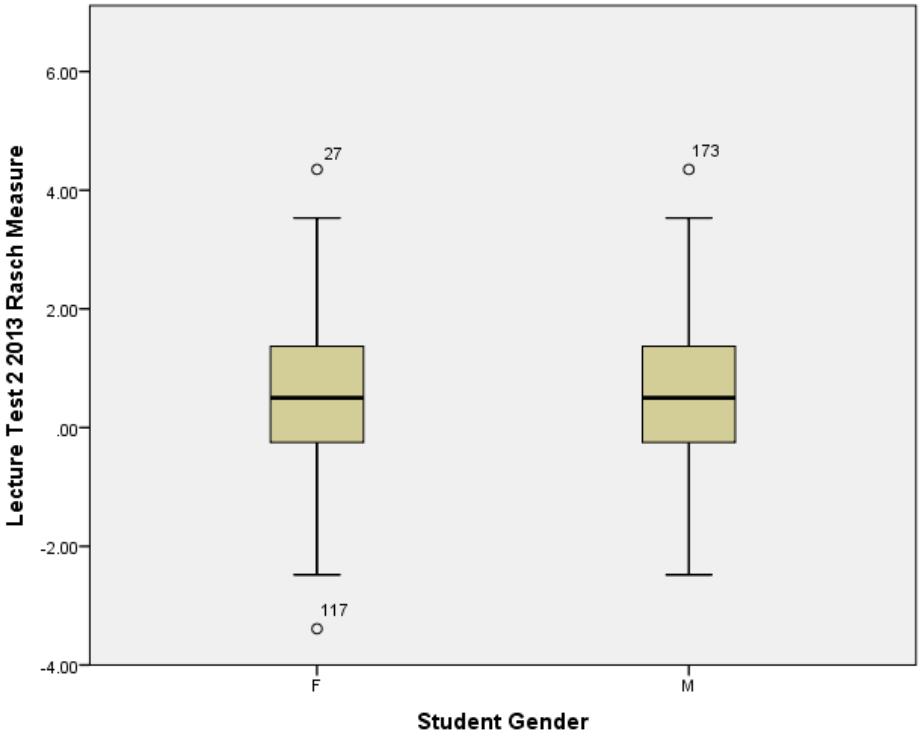

Figure 466: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Lecture Test 2 2013 to Observe Significant Differences

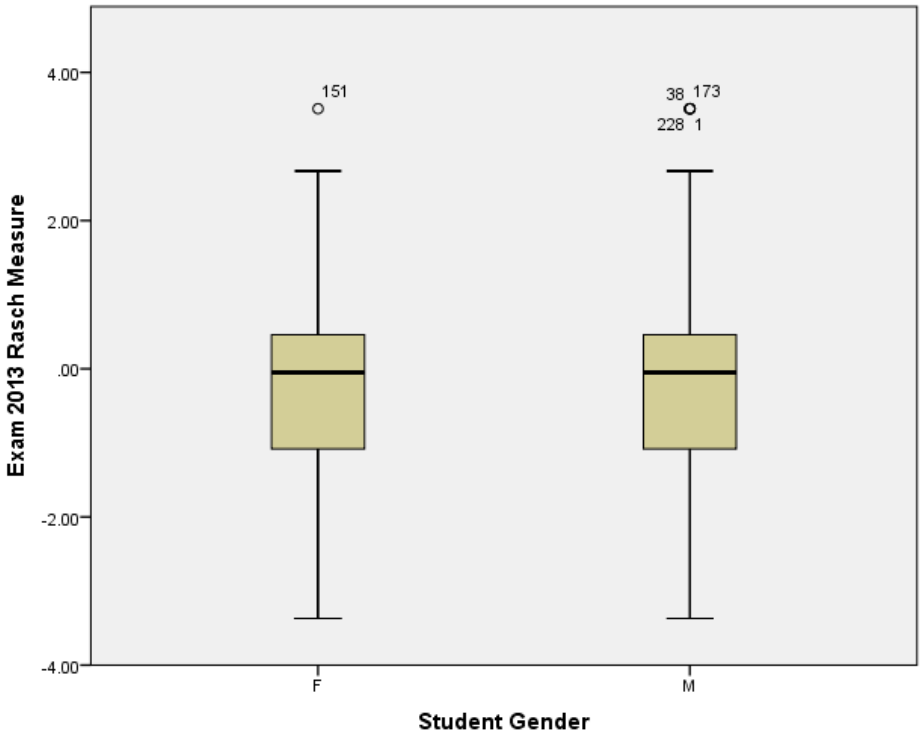

Figure 467: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Exam 2013 to Observe Significant Differences

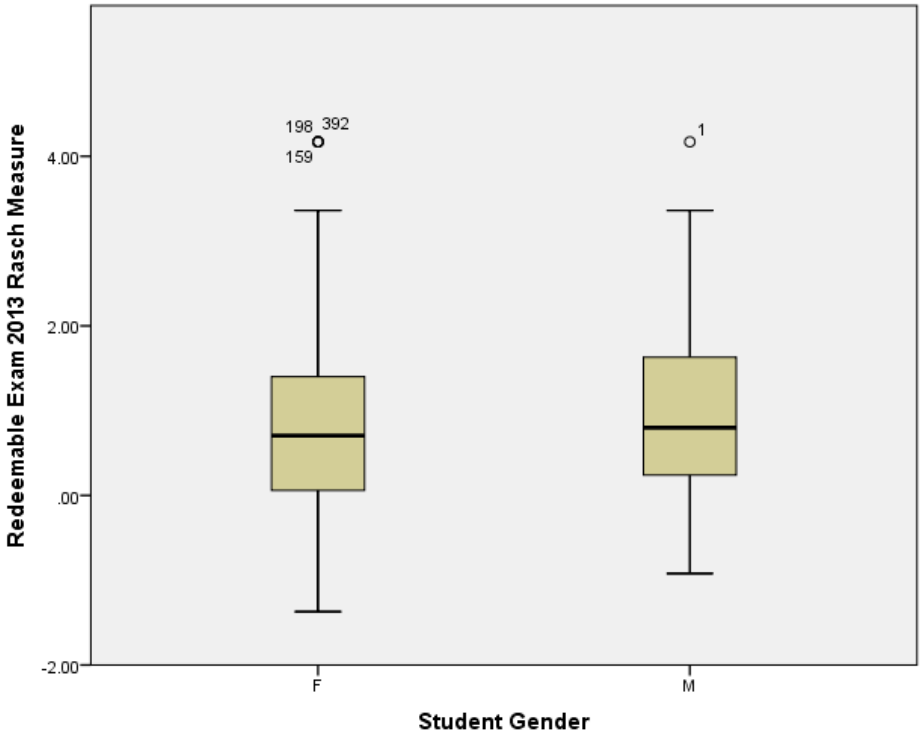

Figure 468: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Redeemable Exam 2013 to Observe Significant Differences

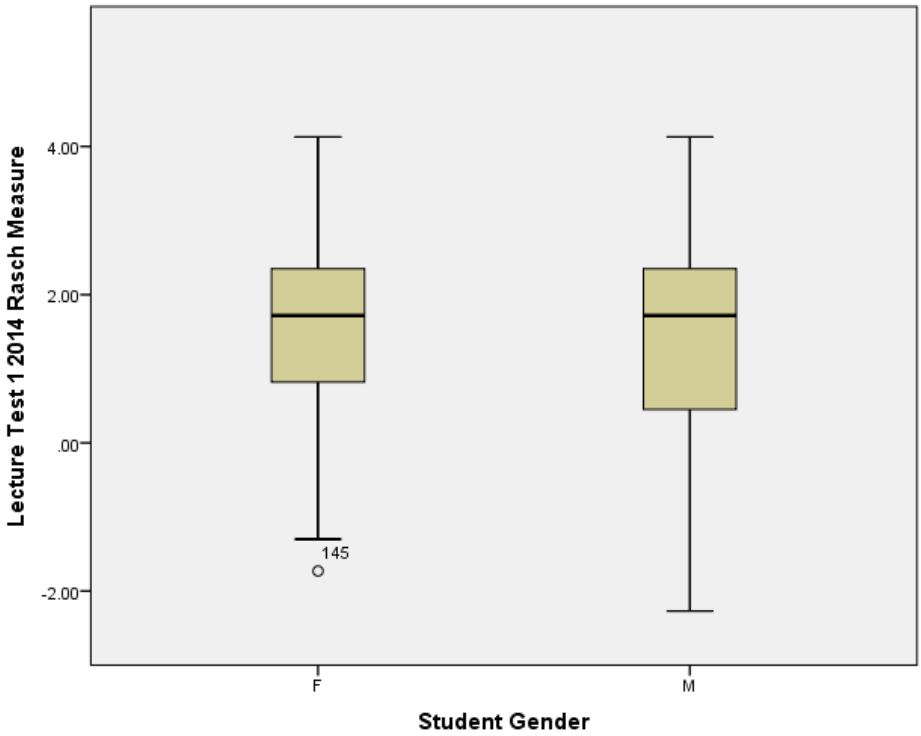

Figure 469: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Lecture Test 1 2014 to Observe Significant Differences

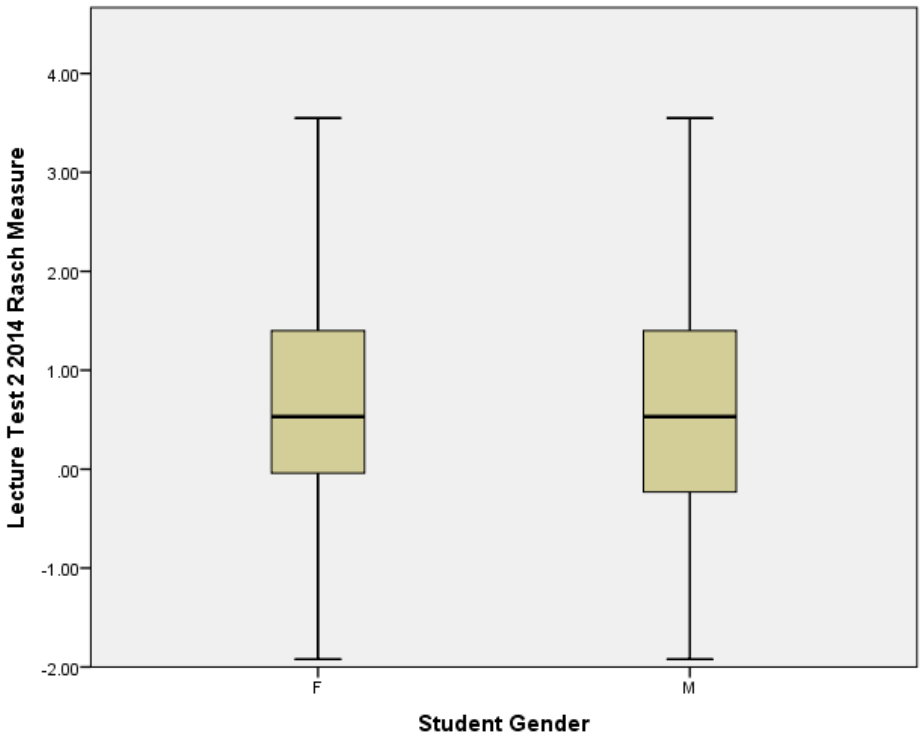

Figure 470: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Lecture Test 2 2014 to Observe Significant Differences

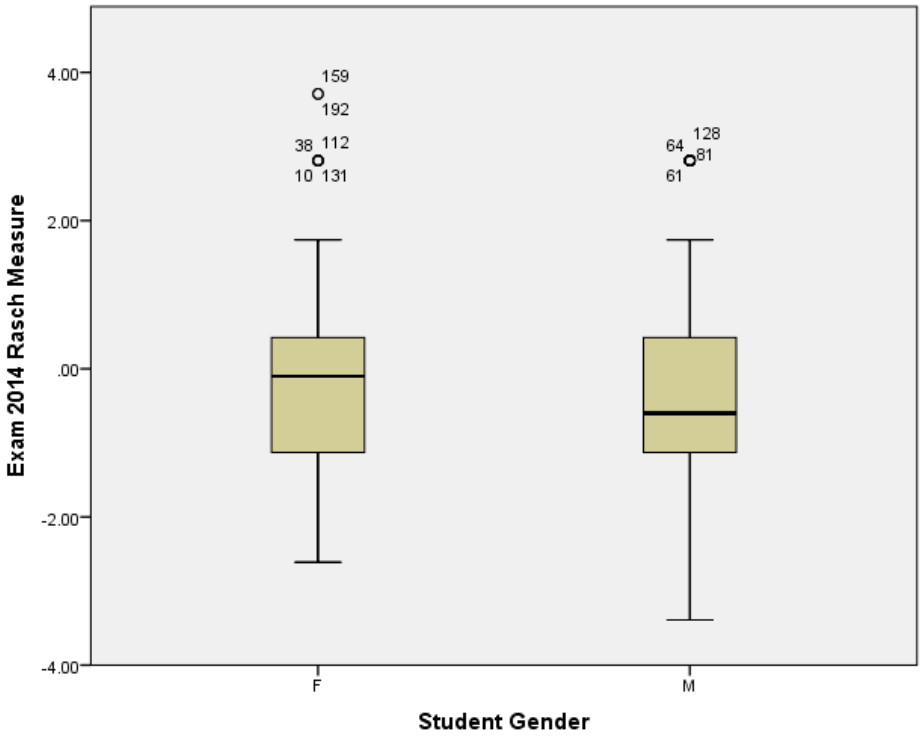

Figure 471: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Exam 2014 to Observe Significant Differences

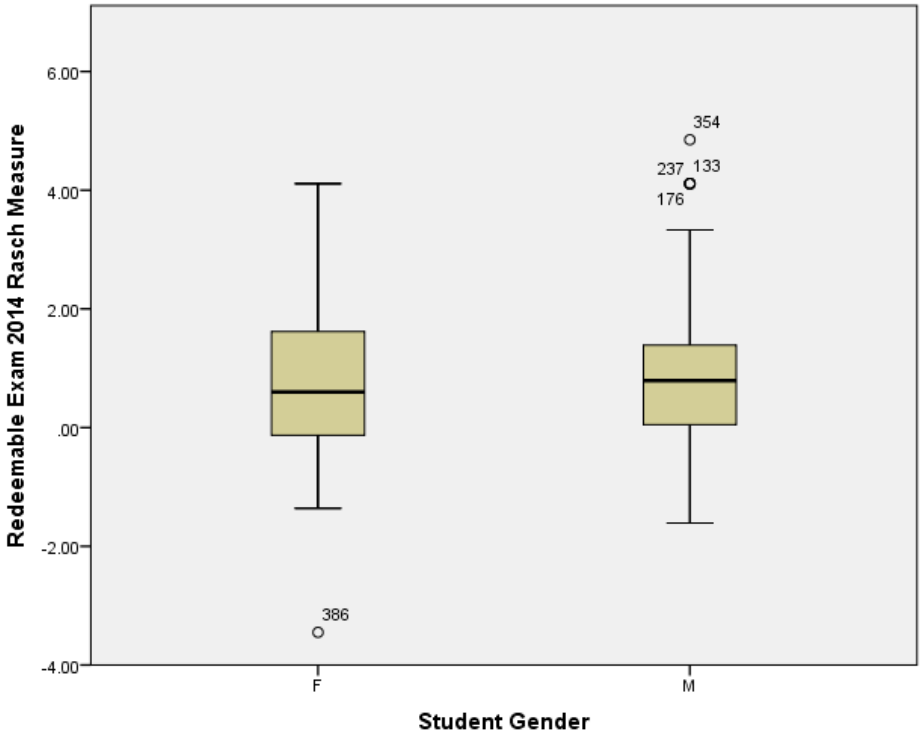

Figure 472: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Redeemable Exam 2014 to Observe Significant Differences

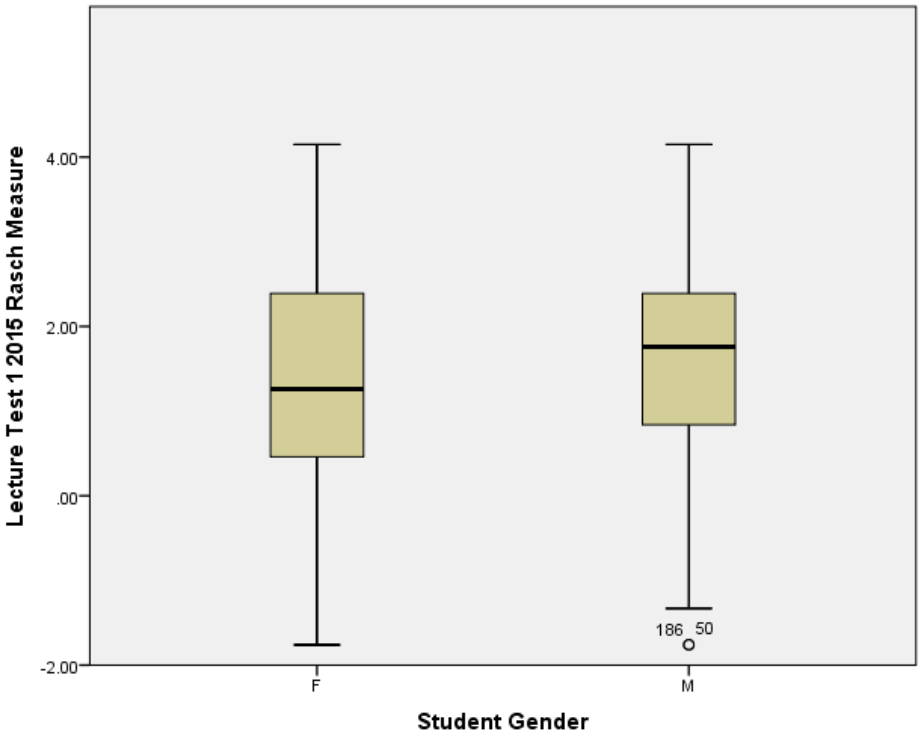

Figure 473: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Lecture Test 1 2015 to Observe Significant Differences

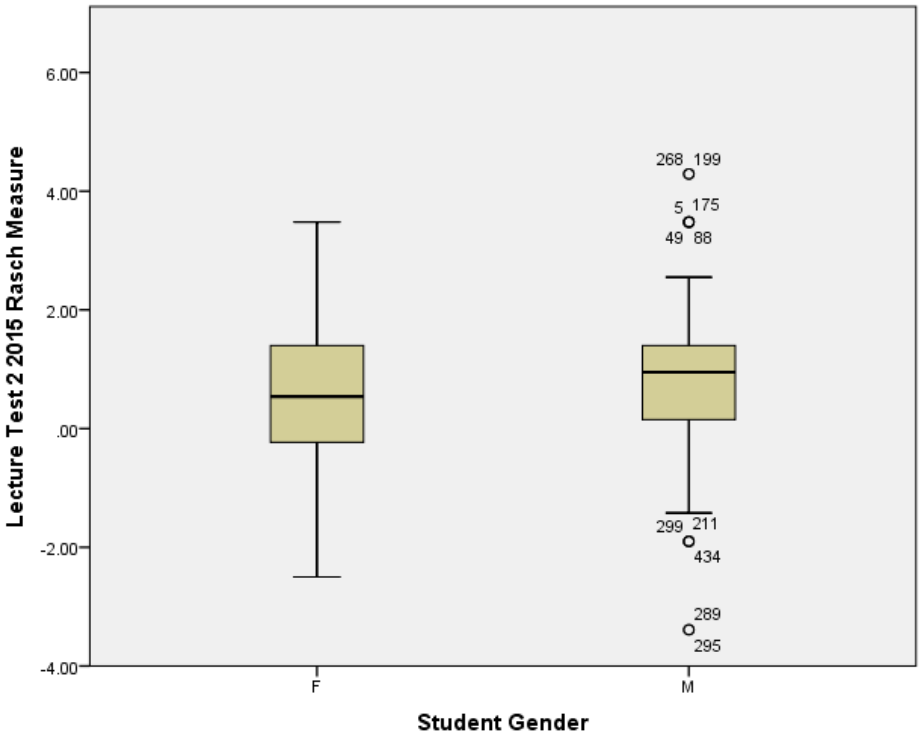

Figure 474: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Lecture Test 2 2015 to Observe Significant Differences

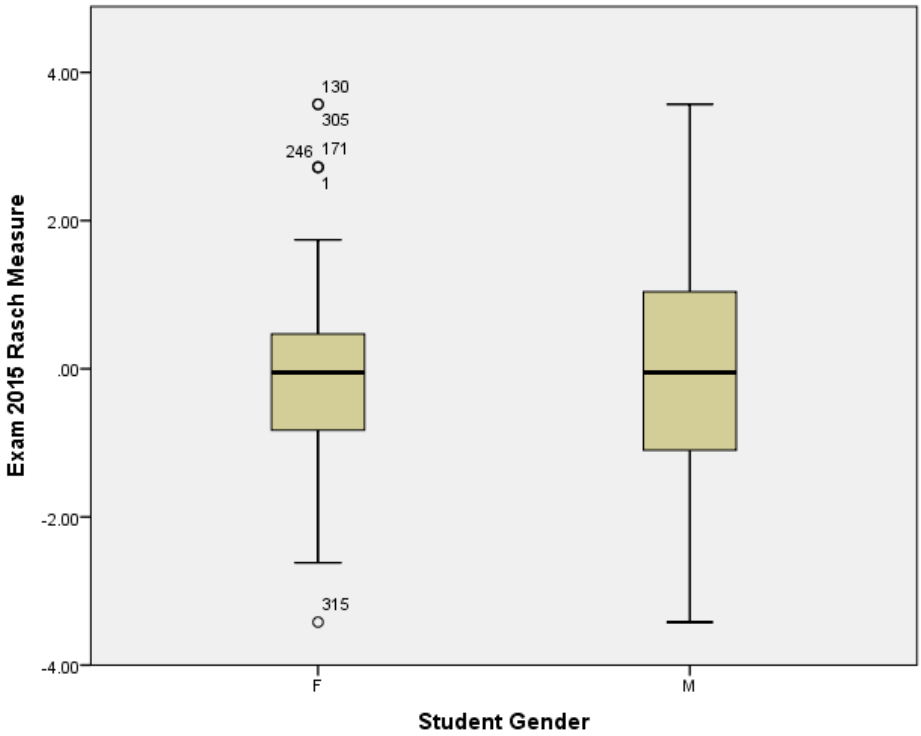

Figure 475: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Exam 2015 to Observe Significant Differences

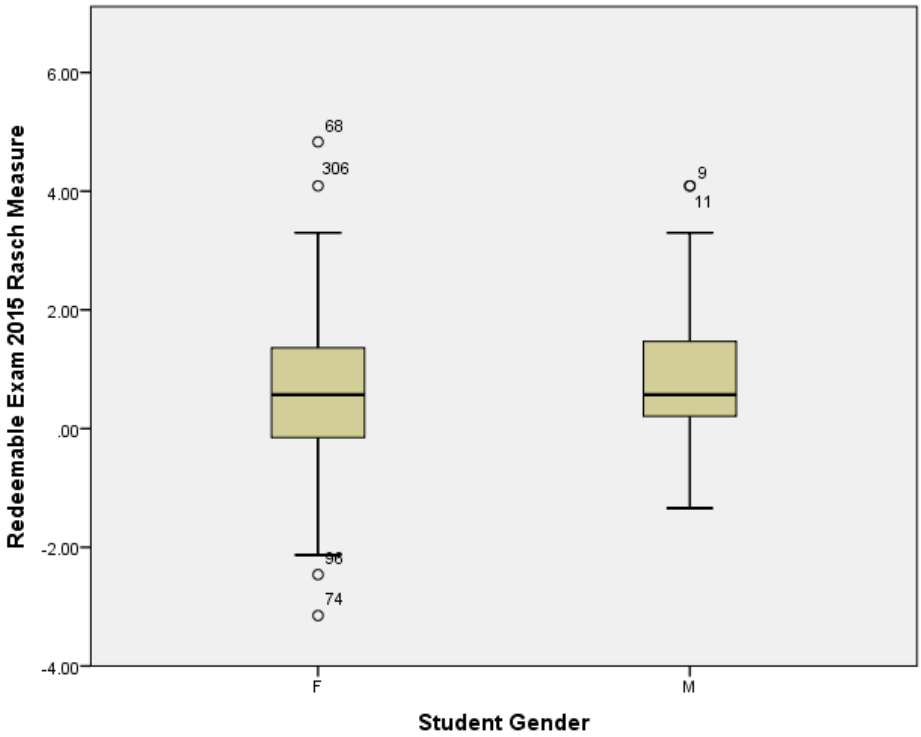

Figure 476: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IA Redeemable Exam 2015 to Observe Significant Differences

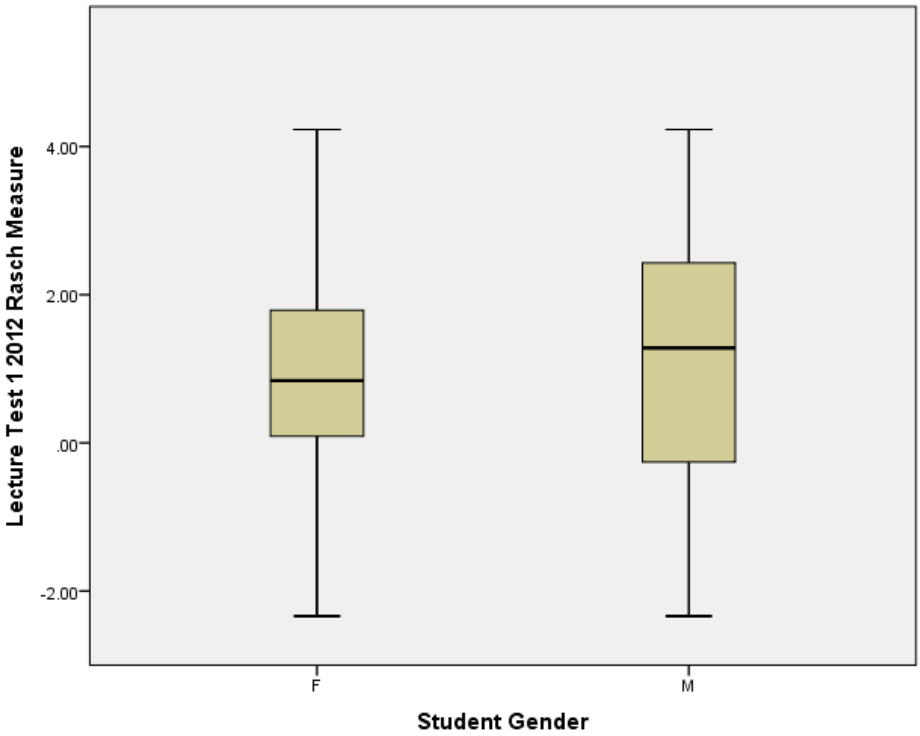

Figure 477: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Lecture Test 1 2012 to Observe Significant Differences

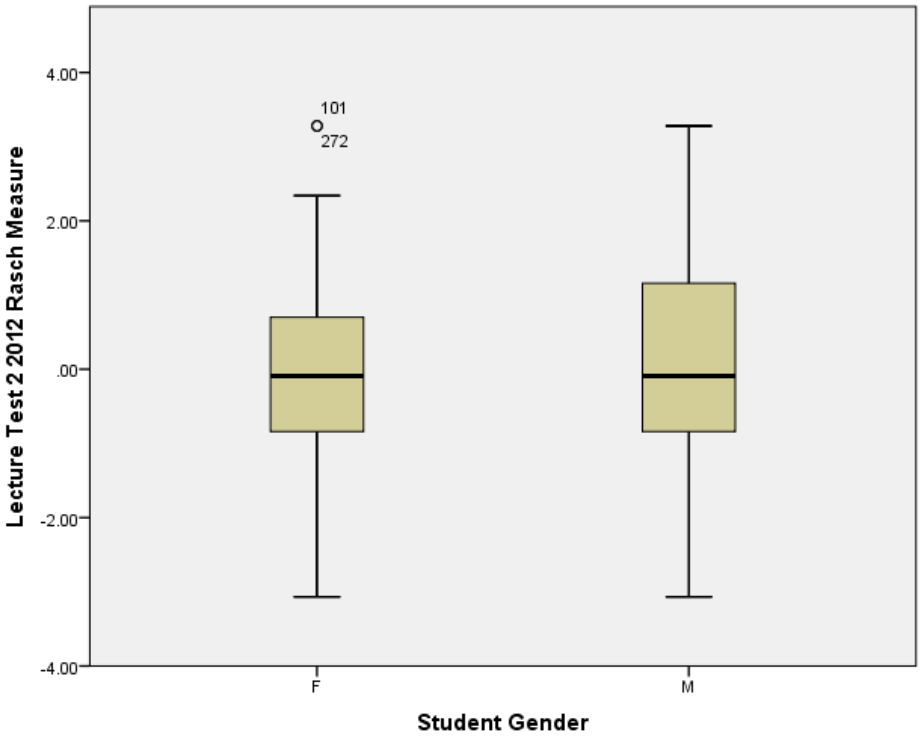

Figure 478: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Lecture Test 2 2012 to Observe Significant Differences

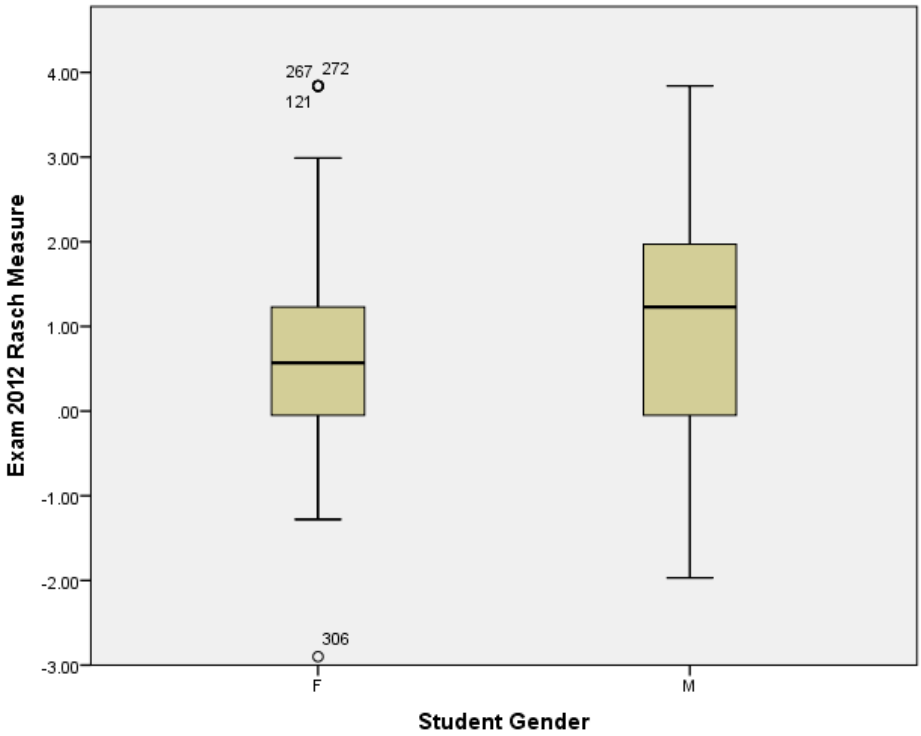

Figure 479: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Exam 2012 to Observe Significant Differences

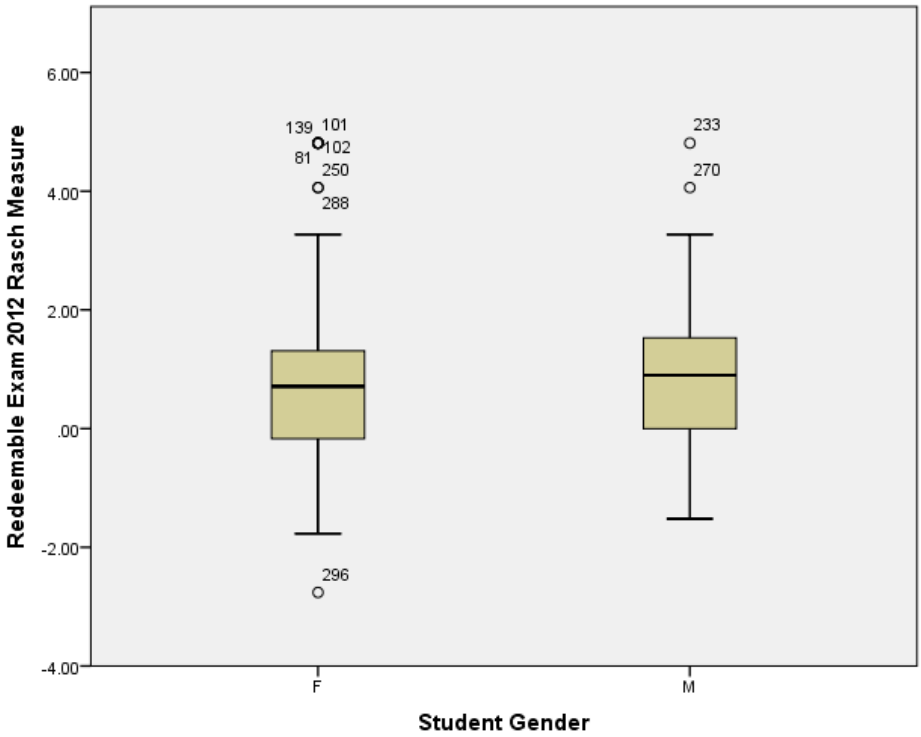

Figure 480: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Redeemable Exam 2012 to Observe Significant Differences

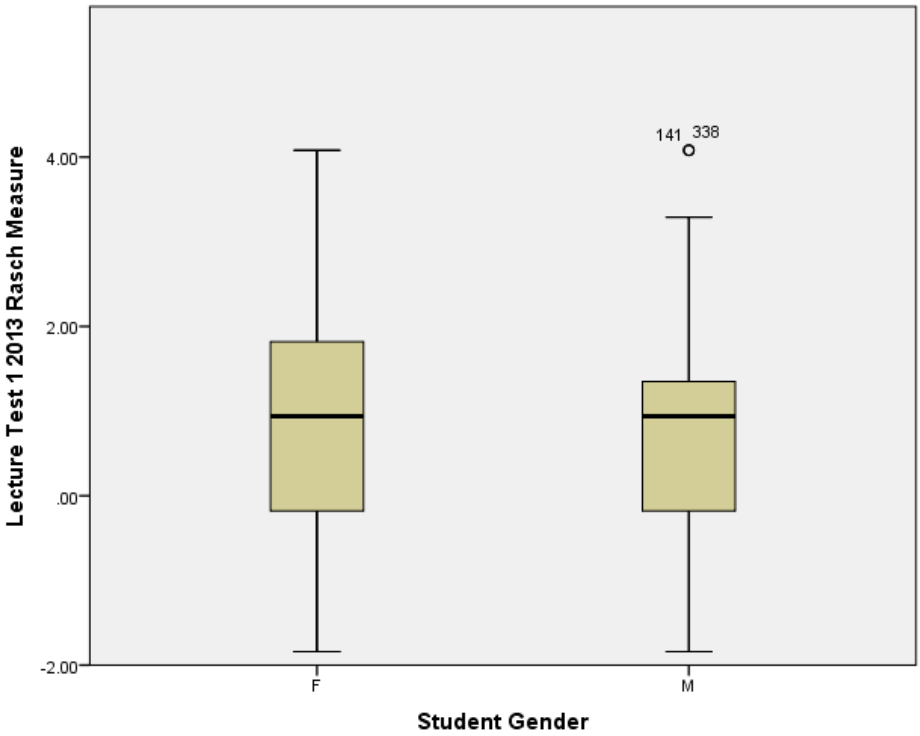

Figure 481: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Lecture Test 1 2013 to Observe Significant Differences

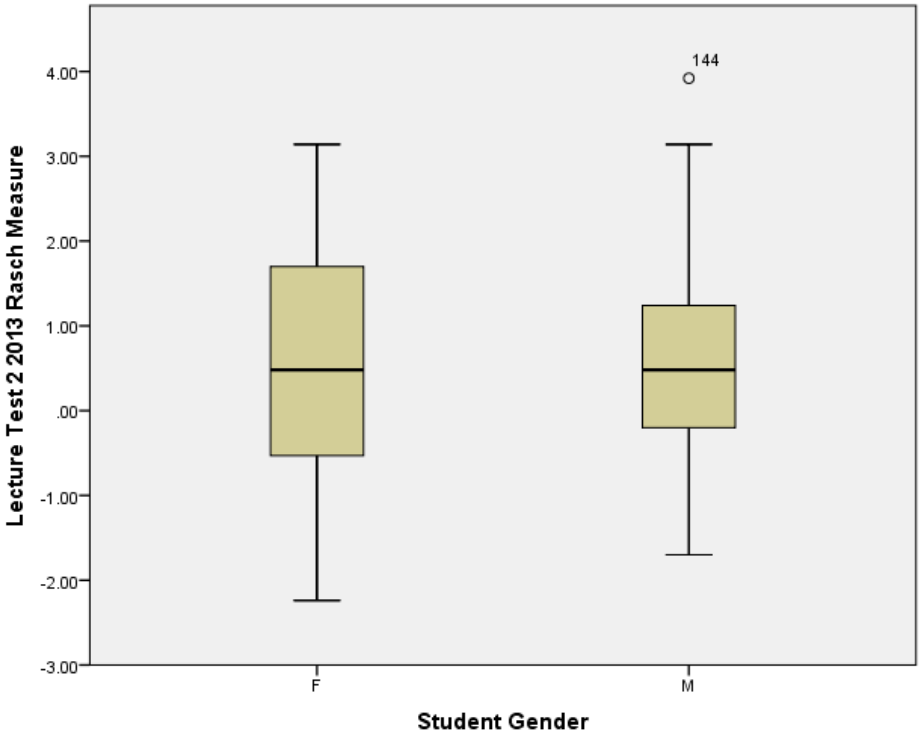

Figure 482: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Lecture Test 2 2013 to Observe Significant Differences

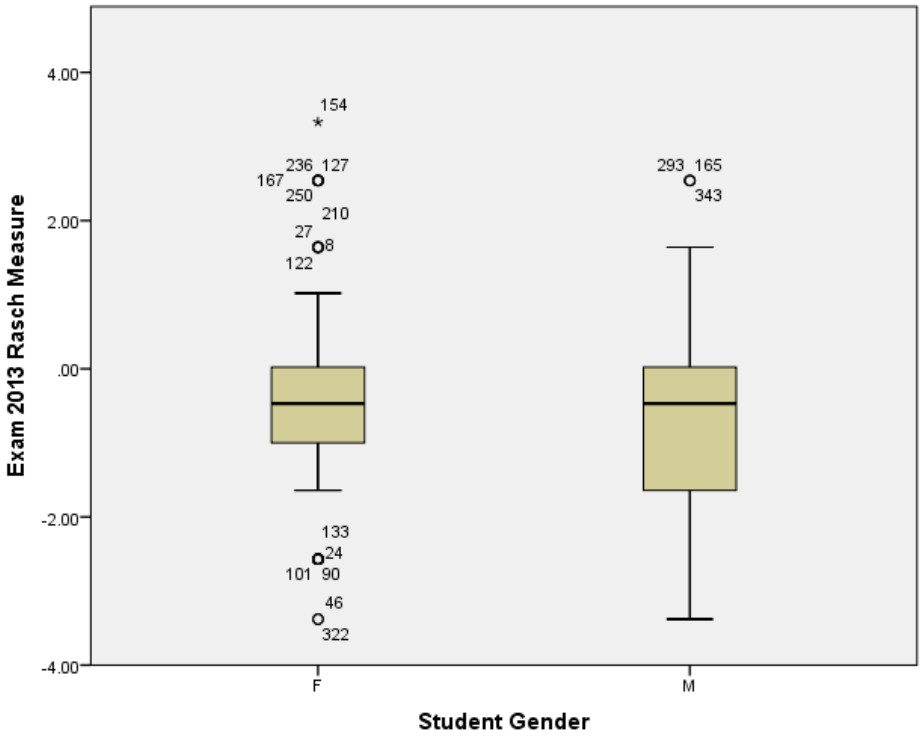

Figure 483: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Exam 2013 to Observe Significant Differences

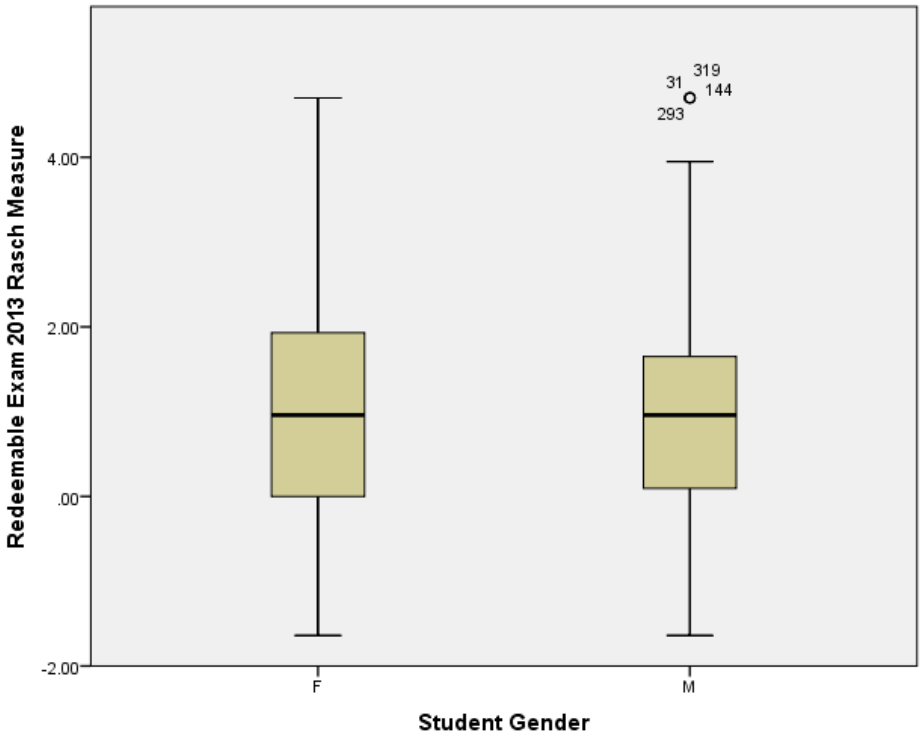

Figure 484: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Redeemable Exam 2013 to Observe Significant Differences

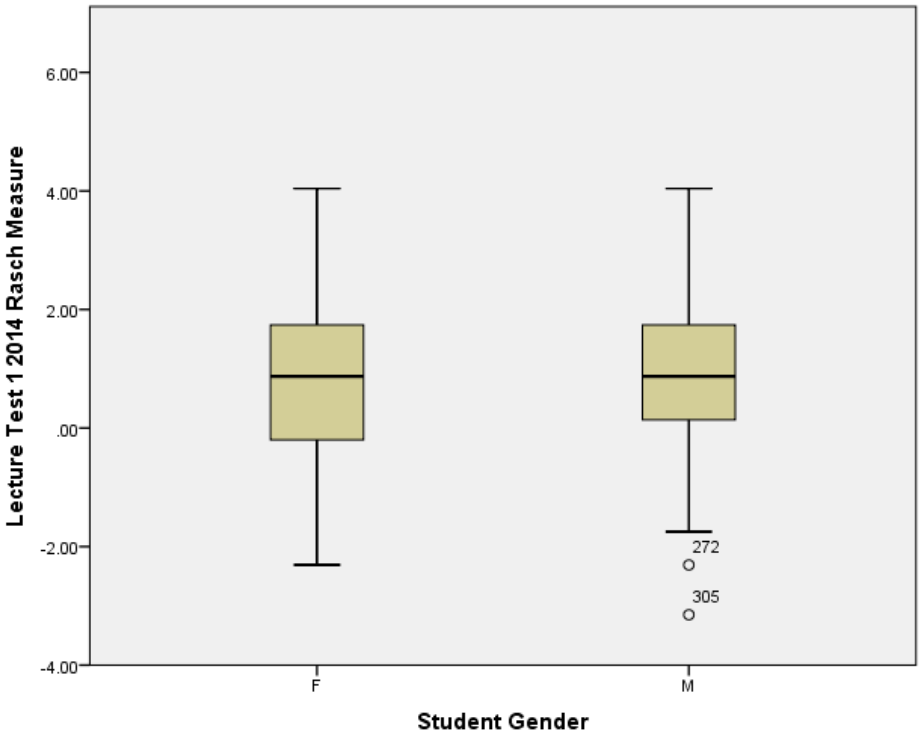

Figure 485: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Lecture Test 1 2014 to Observe Significant Differences

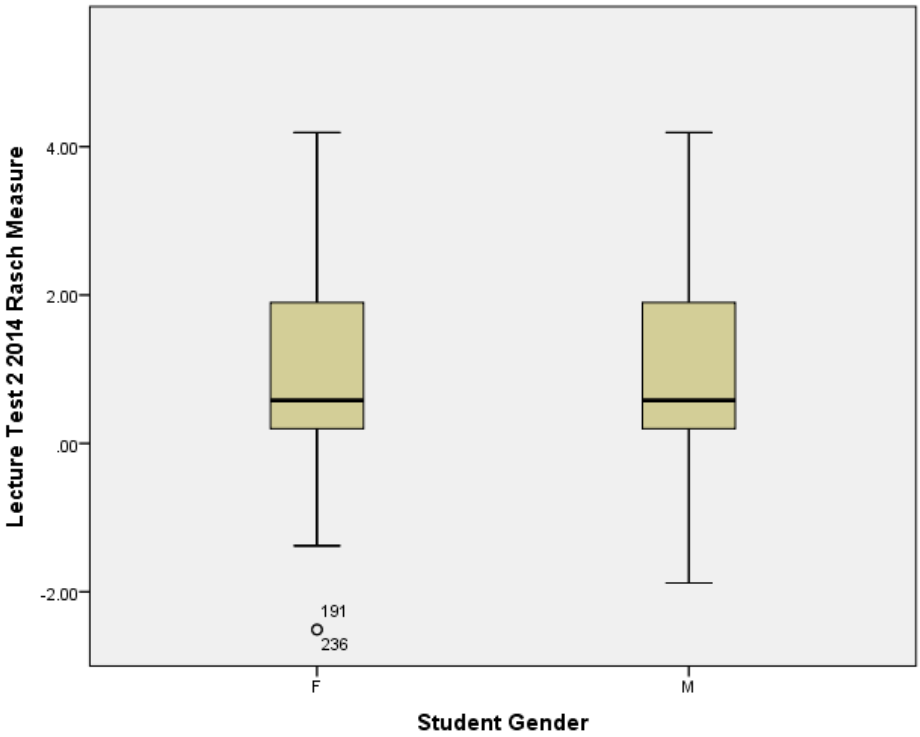

Figure 486: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Lecture Test 2 2014 to Observe Significant Differences

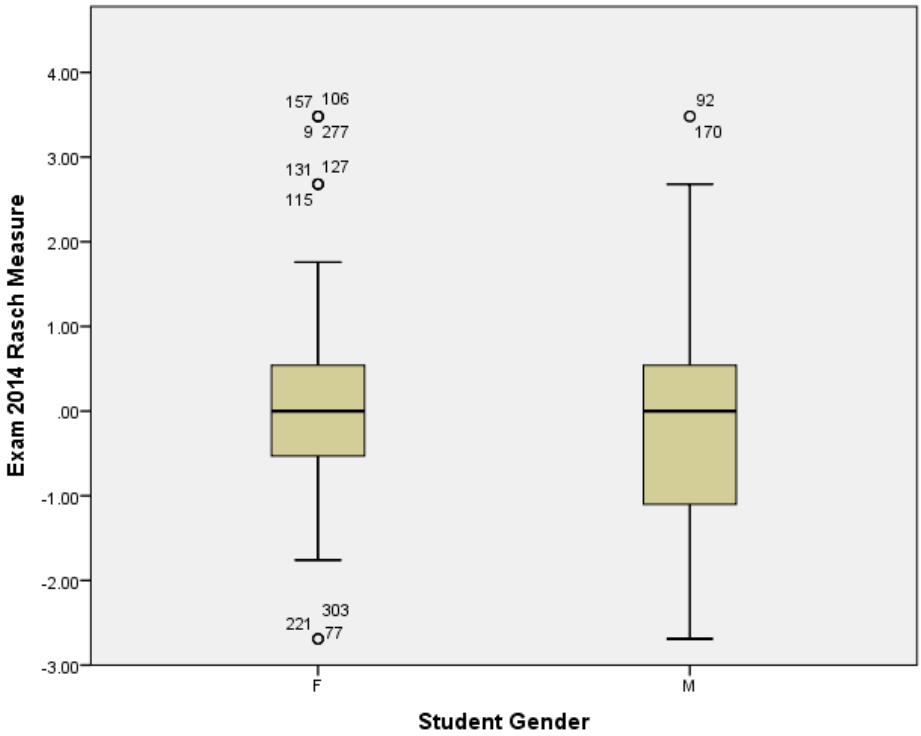

Figure 487: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Exam 2014 to Observe Significant Differences

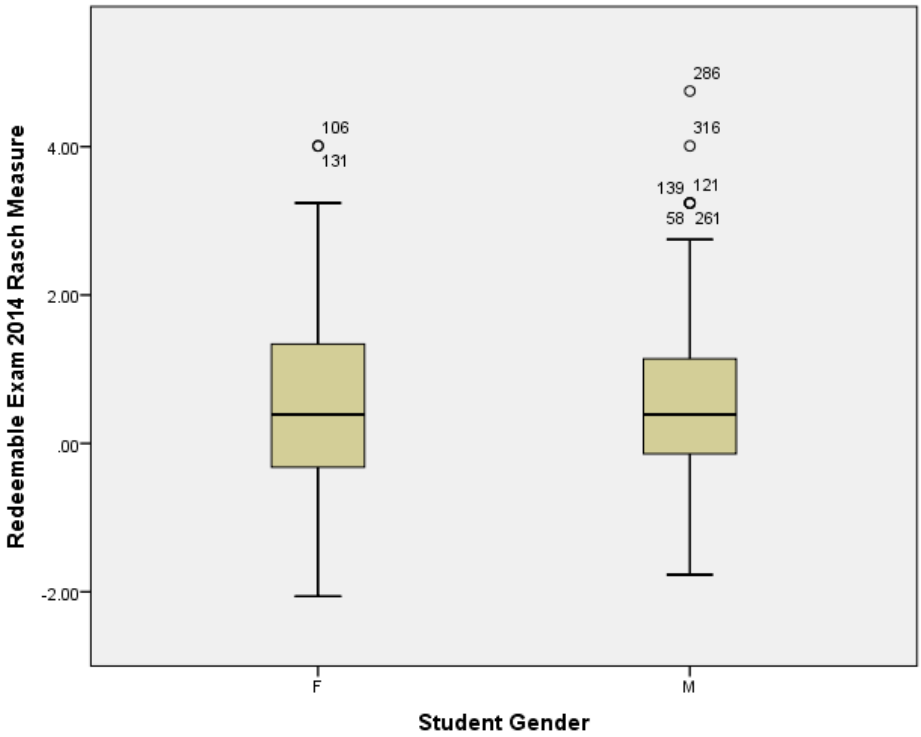

Figure 488: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Redeemable Exam 2014 to Observe Significant Differences

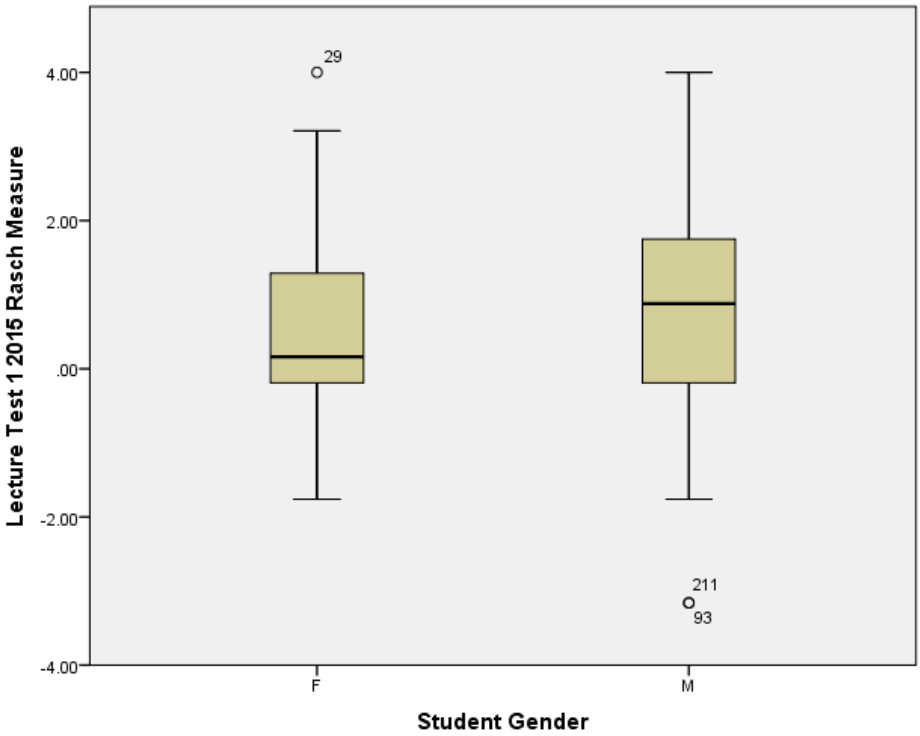

Figure 489: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Lecture Test 1 2015 to Observe Significant Differences

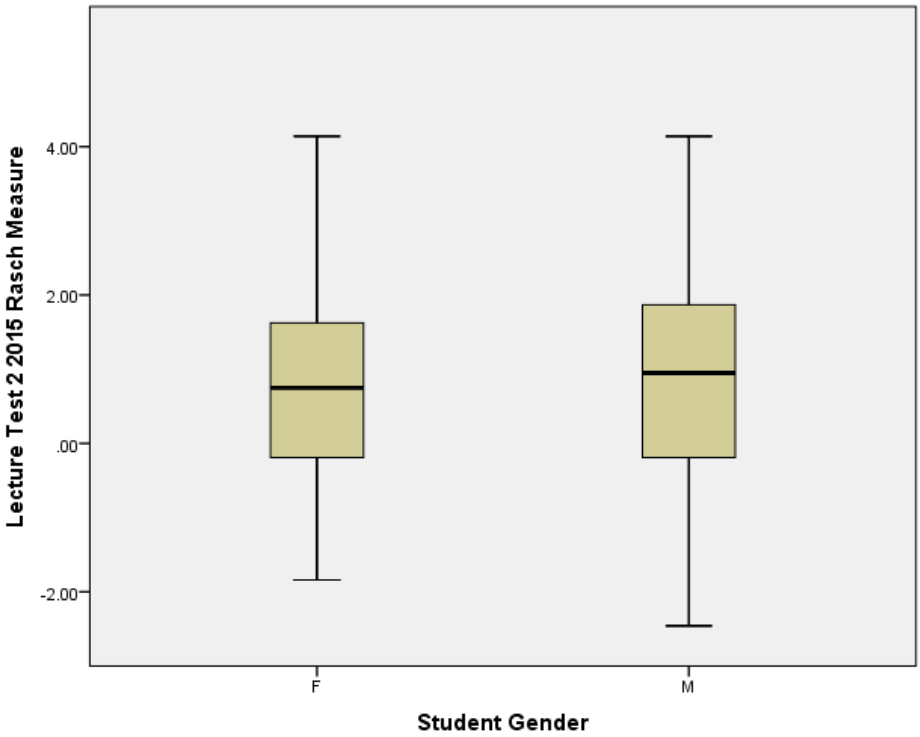

Figure 490: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Lecture Test 2 2015 to Observe Significant Differences

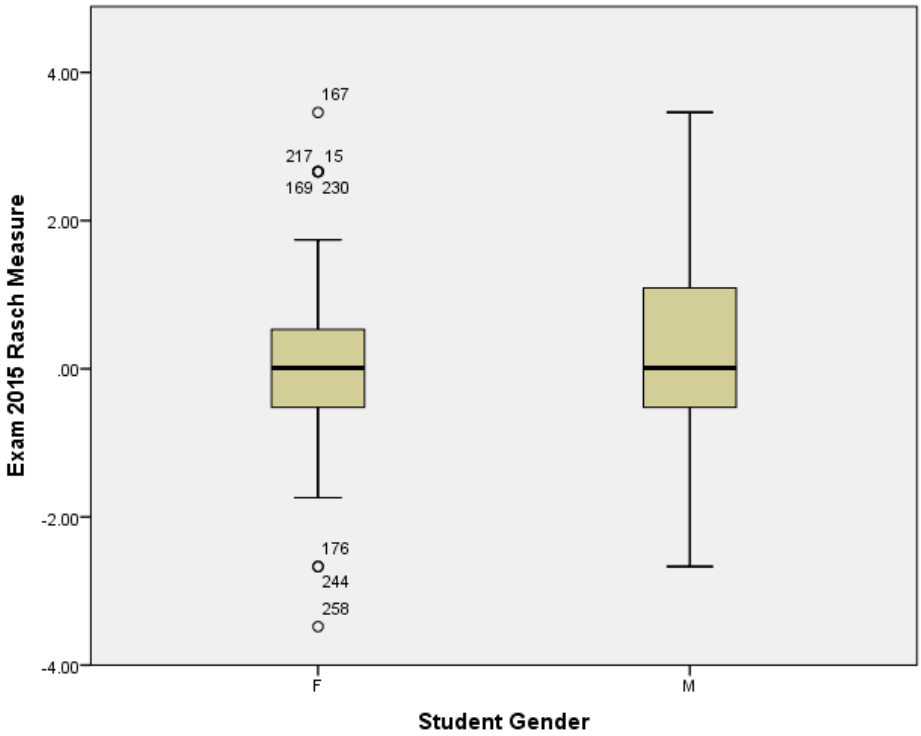

Figure 491: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Exam 2015 to Observe Significant Differences

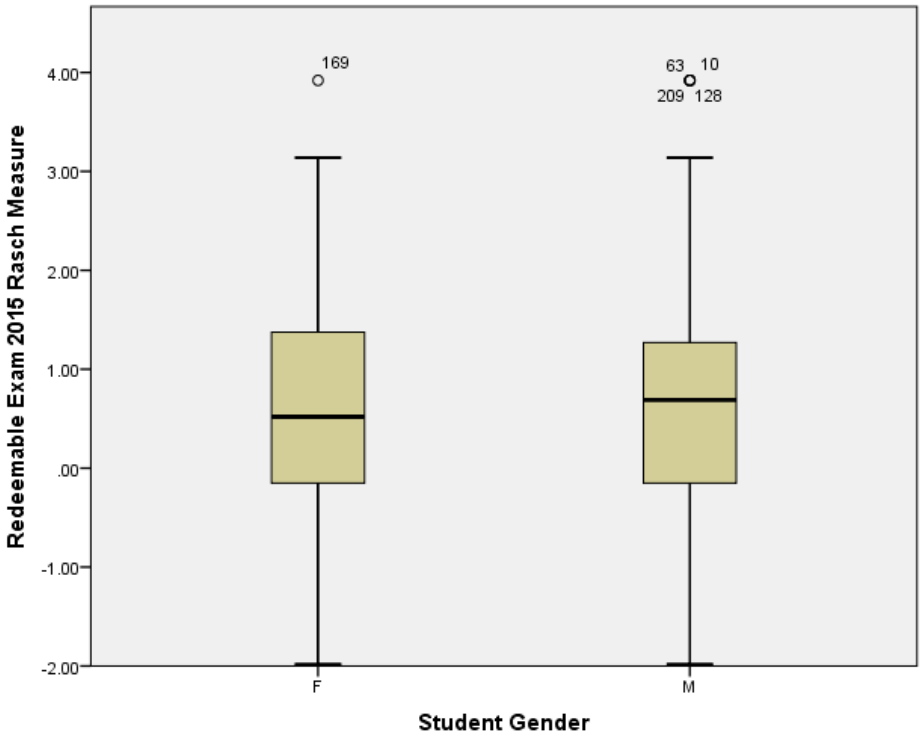

Figure 492: The Boxplot Comparison of Male and Female Student Ability in Foundations of Chemistry IB Redeemable Exam 2015 to Observe Significant Differences

7.15 Gender Bias Items Identified Using of Rasch Analysis

Table 84: Items that were identified to show Significant Differences in the Performance of Male and Female Students thought the use of Rasch Analysis within First-Year Chemistry MCQ Assessment Tasks undertaken at The University of Adelaide

	Item	Times Asked	Year	Item	Gender	ΔDIF	*t*-value
Chemistry IA	Lec_1_8	4	2014 - 2015				
			2014	Exam_2_8	Male	0.72	3.53
			2015	Exam_2_8	Male	0.42	2.17
	Lec_1_9	4	2014-2015				
			2014	Exam_2_9	Male	0.52	2.35
			2015	Exam_2_9	Male	0.47	2.23
	Lec_1_15	2	2012-2013				
			2012	Lec_1_15	Female	-0.51	-2.39
			2013	Lec_1_15	Female	-0.54	-2.49
Chemistry IB	Lec_1_3	8	2012-2015				
			2012	Exam_2_3	Male	0.48	2.15
			2013	Exam_2_3	Male	0.45	2.02
	Lec_1_4	8	2012-2015				
			2013	Exam_2_4	Male	0.62	2.74
			2014	Lec_2_4	Male	0.58	2.56
			2014	Exam_2_4	Male	0.46	2.17
			2015	Lec_1_4	Male	0.57	2.46
	Lec_1_5	8	2012-2015				
			2012	Lec_1_5	Female	-0.46	-2.01
			2014	Lec_1_5	Female	-0.51	-2.26
	Lec_1_8	8	2012-2015				
			2012	Lec_1_8	Male	0.84	3.56
			2012	Exam_2_8	Male	0.52	2.25
			2013	Lec_1_8	Male	0.53	2.21
			2013	Exam_2_8	Male	0.48	2.05
	Lec_2_12	8	2012-2015	*Q26 in redeemable exam*			
			2012	Exam_2_26	Female	-0.49	-2.25
			2014	Lec_2_12	Female	-0.47	-2.03
	Lec_2_15	8	2012-2015	*Q29 in redeemable exam*			
			2013	Exam_2_29	Female	-0.6	-2.67
			2015	Lec_2_15	Female	-0.49	-2.16
	Exam_1_9	4	2012-2015				
			2012	Exam_1_9	Male	0.47	2.07
			2015	Exam_1_9	Male	0.45	2.13
Foundations of	Lec_1_7	4	2012-2015				
			2013	Lec_1_7	Male	1.09	3.69
			2015	Lec_1_7	Male	0.65	2.2

	Lec_1_14	1	2012				
			2012	Lec_1_14	Female	-0.77	-2.44
Foundations of Chemistry IB	**Lec_1_14 [2012: Lec_2_5]**	4	2012-2015				
			2013	Lec_1_14	Female	-0.71	-2.24
			2014	Lec_1_14	Male	0.77	2.05
	Red_Exam_11 [2012: Red_Exam_17]	4	2012-2015	*Variation to Lec_1_11 [2012: Lec_2_2]*			
			2013	Exam_2_11	Female	-0.61	-2.21
			2014	Exam_2_11	Female	-0.69	-2.42

7.16 Item Breakdown Histograms using MCQ Classification Process

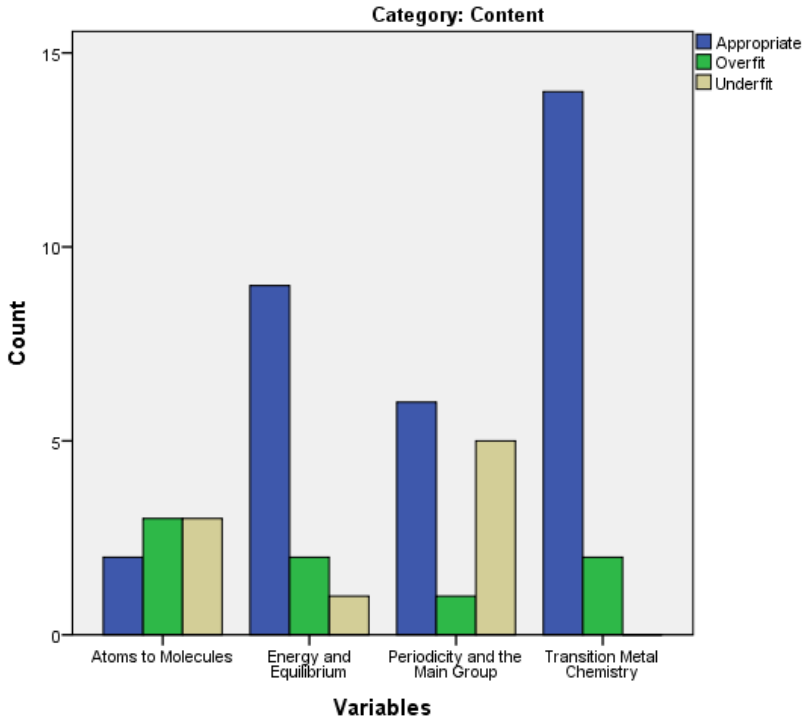

Figure 493: The Number of Unique Items Present for Each Topic Covered within Chemistry IA from 2012-2015

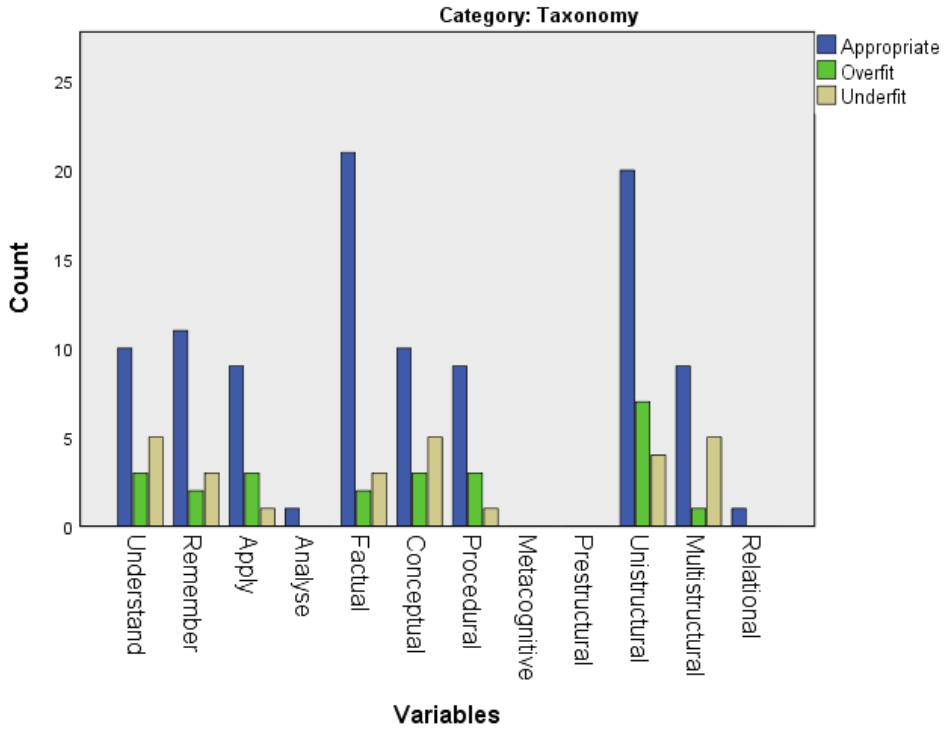

Figure 494: The Number of Unique Items Present for Each Taxonomy within Chemistry IA from 2012-2015

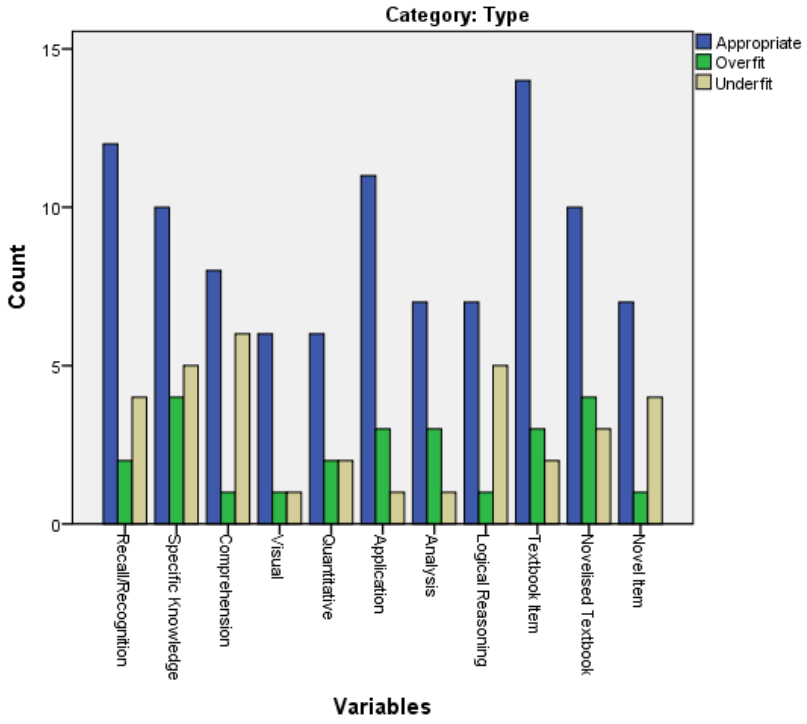

Figure 495: The Number of Unique Items Present for Each Item Type within Chemistry IA from 2012-2015

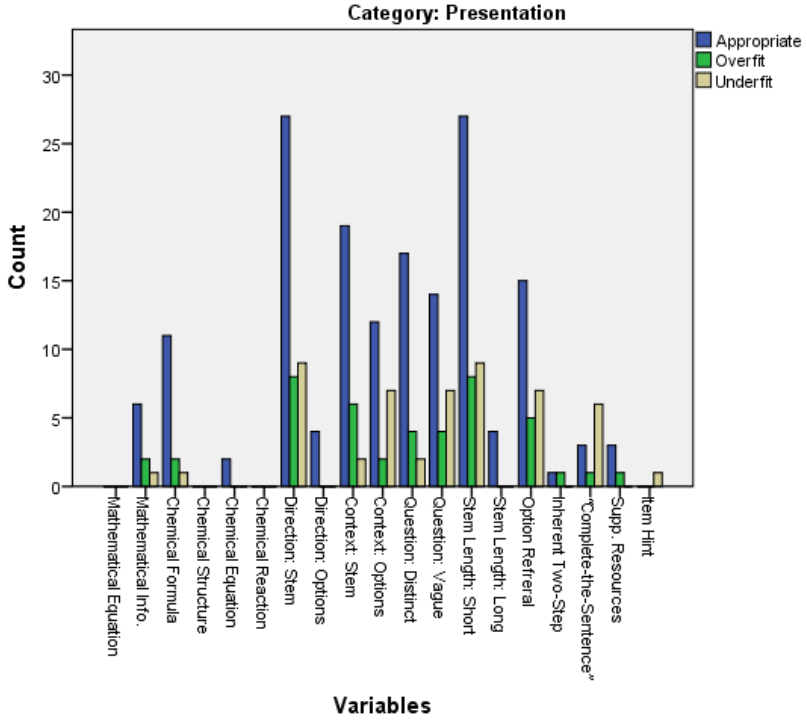

Figure 496: The Number of Unique Items Present for Each Item Presentation Style within Chemistry IA from 2012-2015

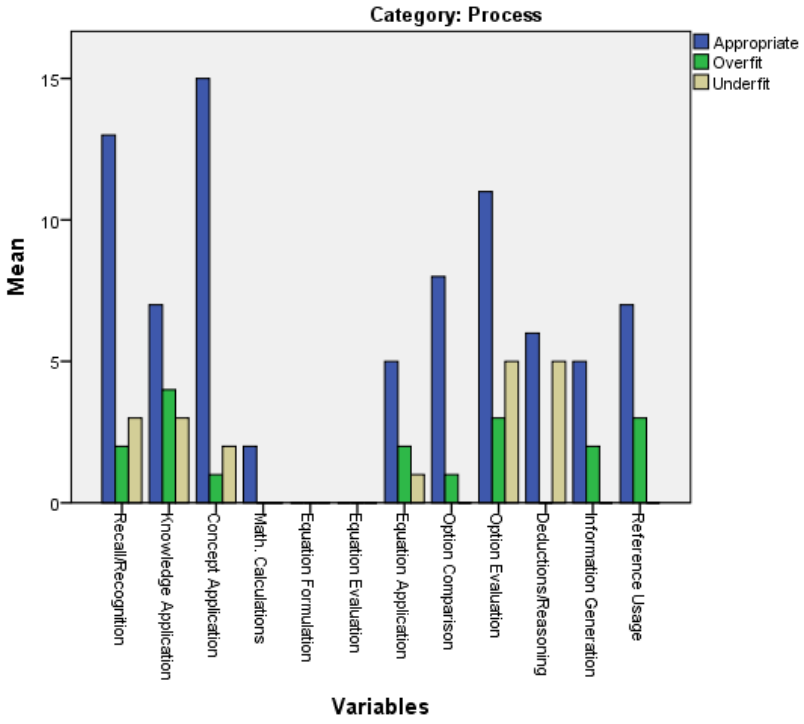

Figure 497: The Number of Unique Items Present for Each Item Process within Chemistry IA from 2012-2015

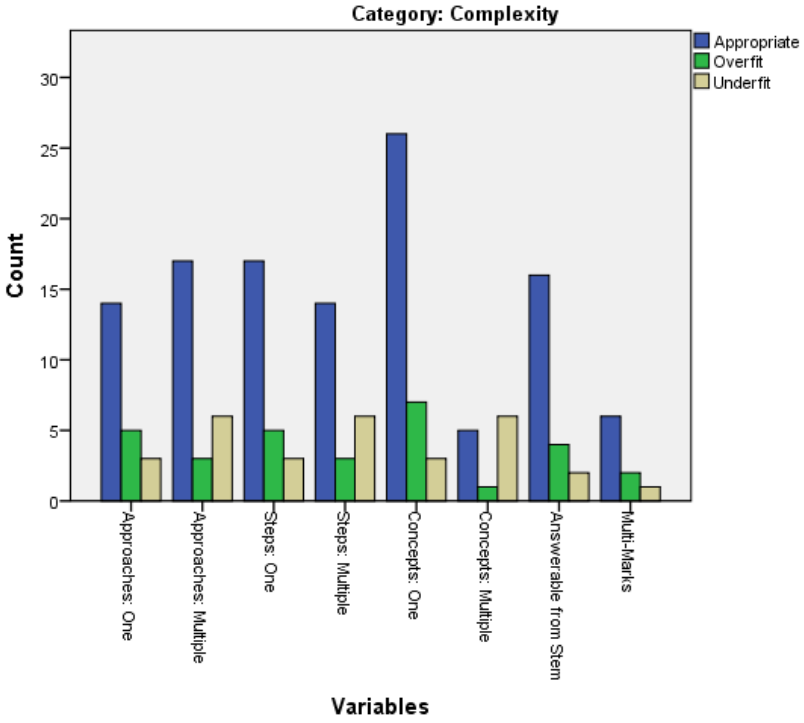

Figure 498: The Number of Unique Items Present for Each Complexity Level within Chemistry IA from 2012-2015

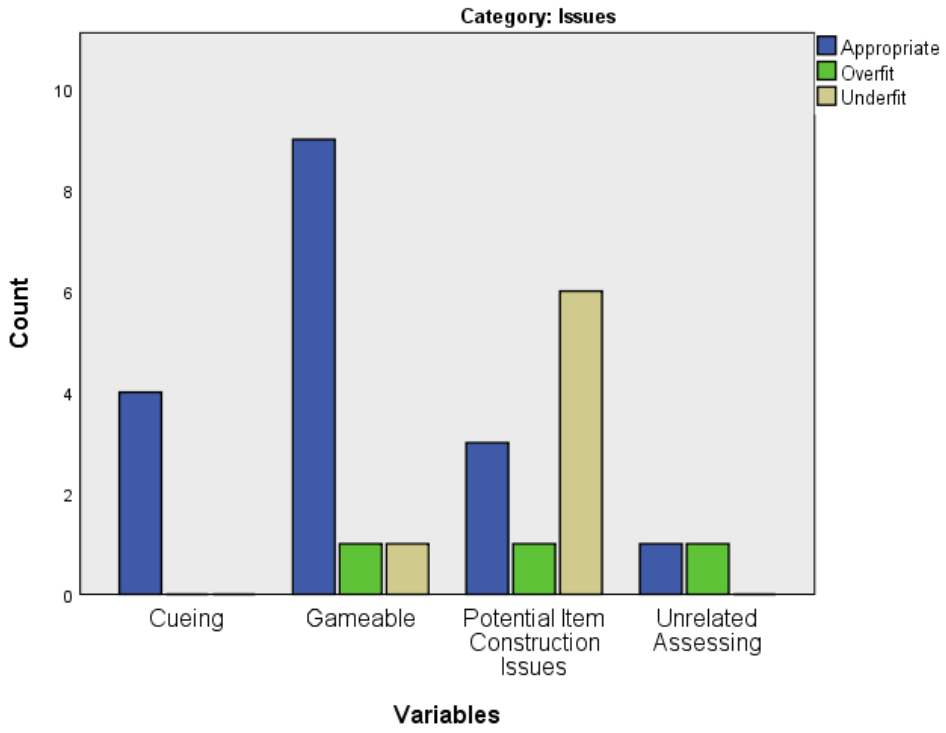

Figure 499: The Number of Unique Items Present with a Potential Item Flaws within Chemistry IA from 2012-2015

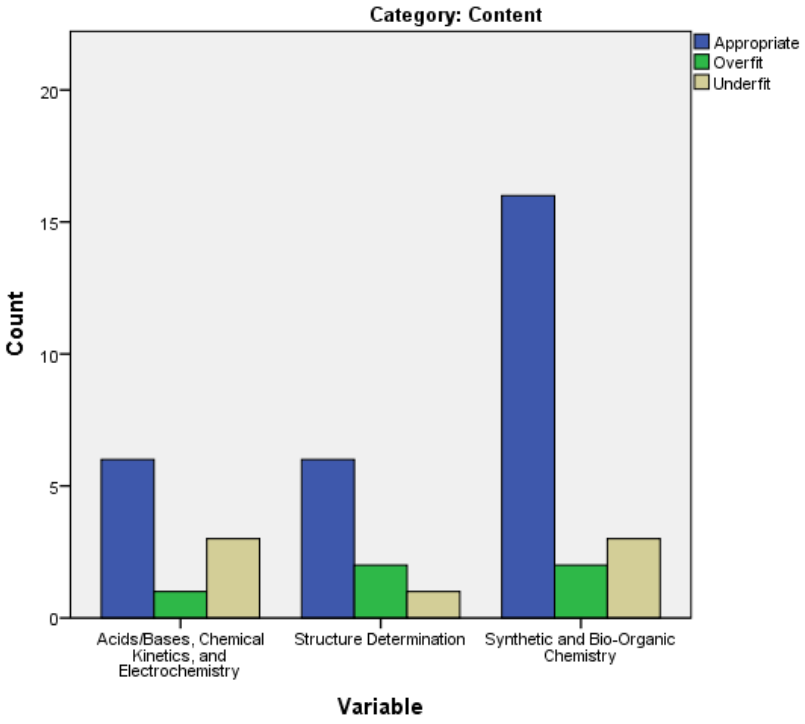

Figure 500: The Number of Unique Items Present for Each Topic Covered within Chemistry IB from 2012-2015

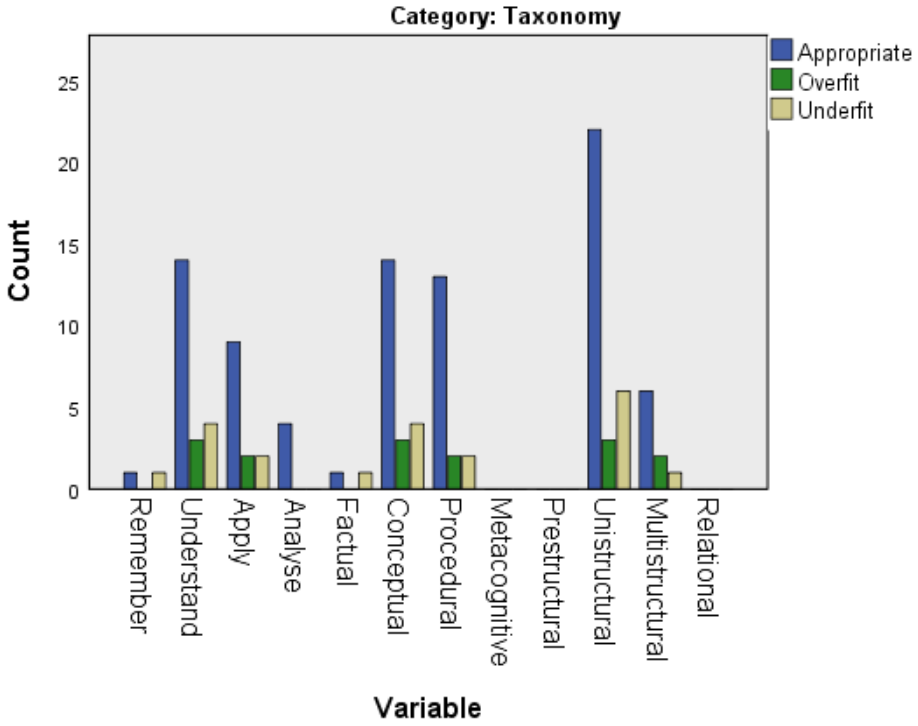

Figure 501: The Number of Unique Items Present for Each Taxonomy within Chemistry IB from 2012-2015

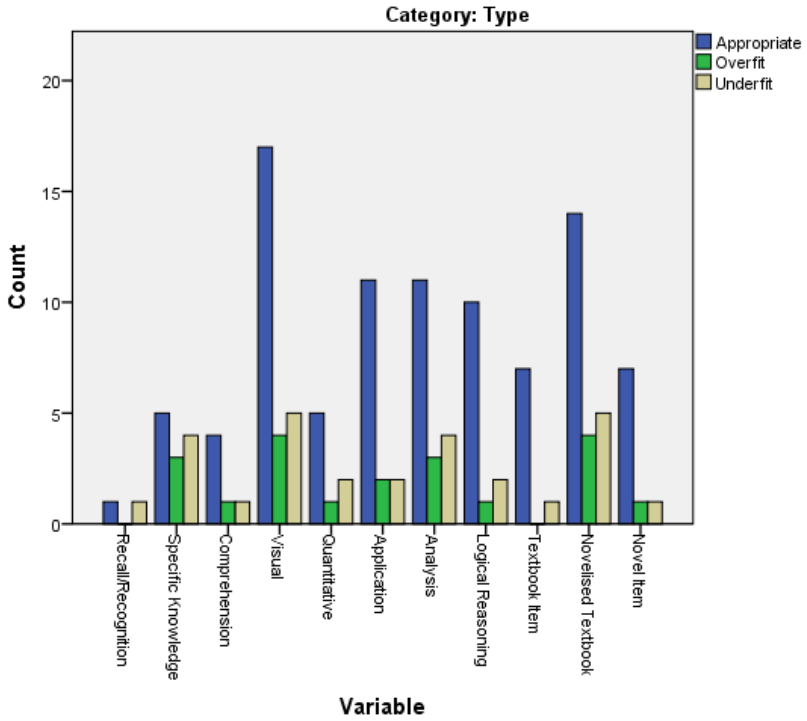

Figure 502: The Number of Unique Items Present for Each Item Type within Chemistry IB from 2012-2015

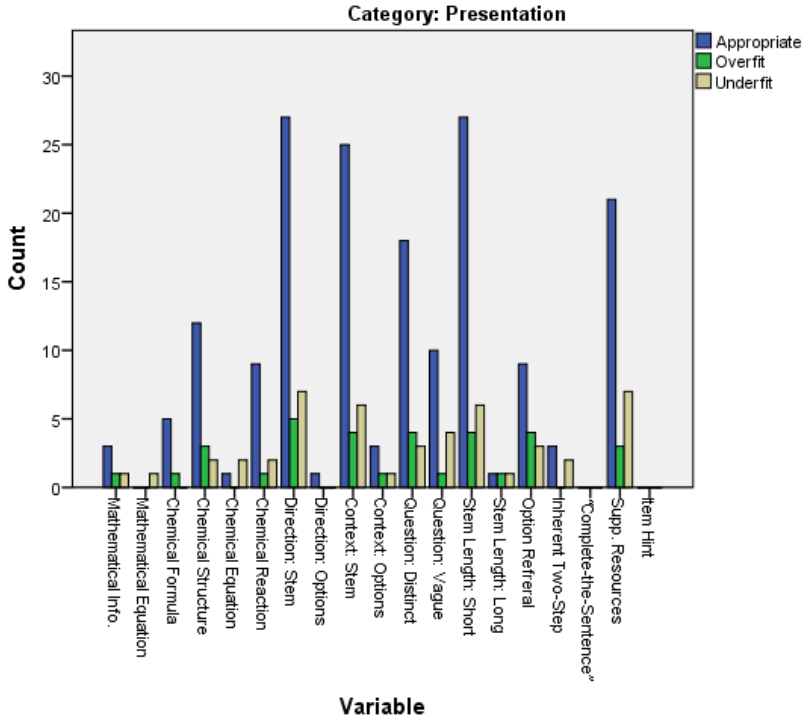

Figure 503: The Number of Unique Items Present for Each Item Presentation Style within Chemistry IB from 2012-2015

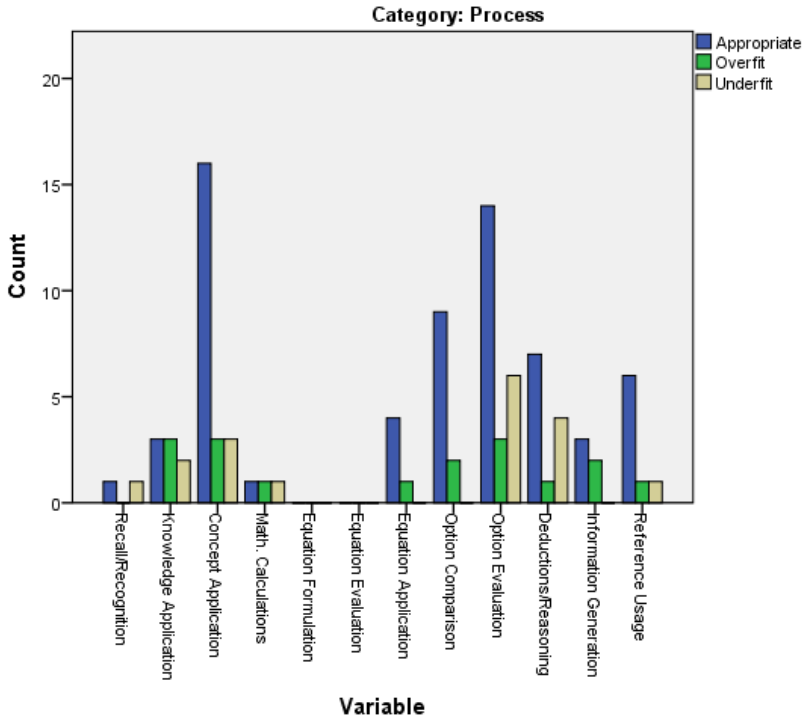

Figure 504: The Number of Unique Items Present for Each Item Process within Chemistry IB from 2012-2015

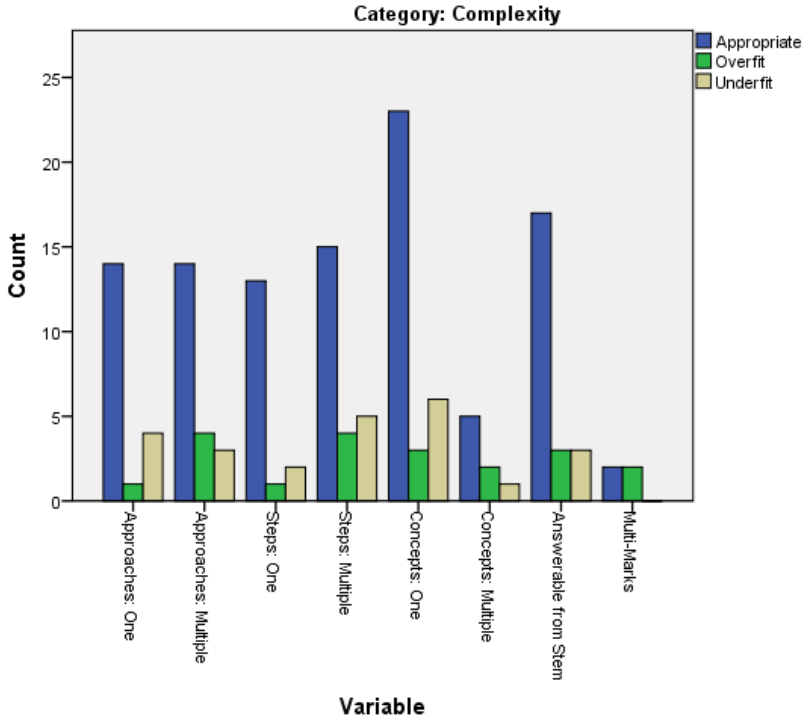

Figure 505: The Number of Unique Items Present for Each Level of Item Complexity within Chemistry IB from 2012-2015

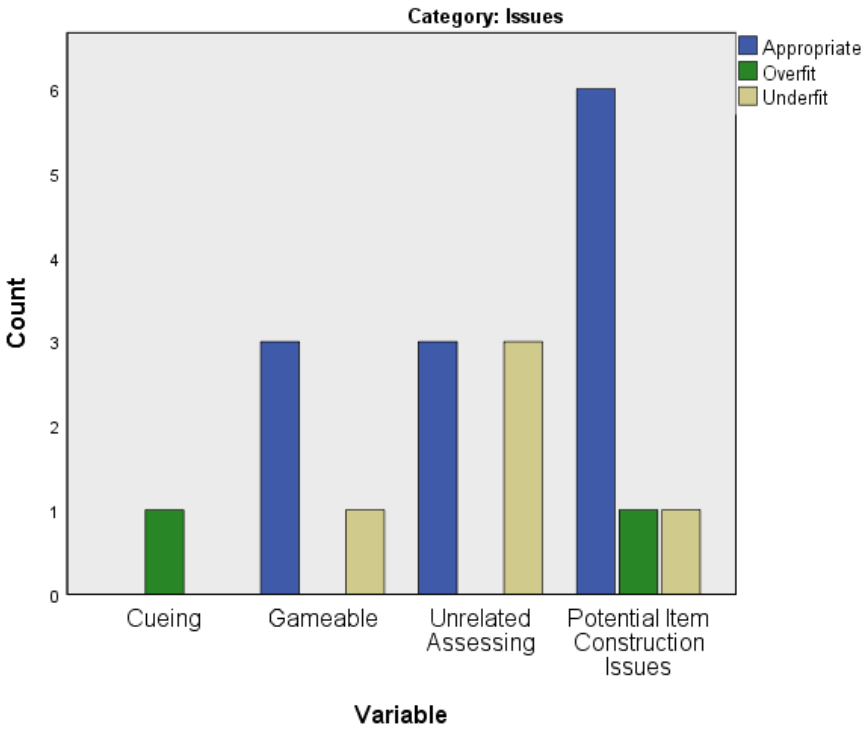

Figure 506: The Number of Unique Items Present with a Potential Item Flaws within Chemistry IB from 2012-2015

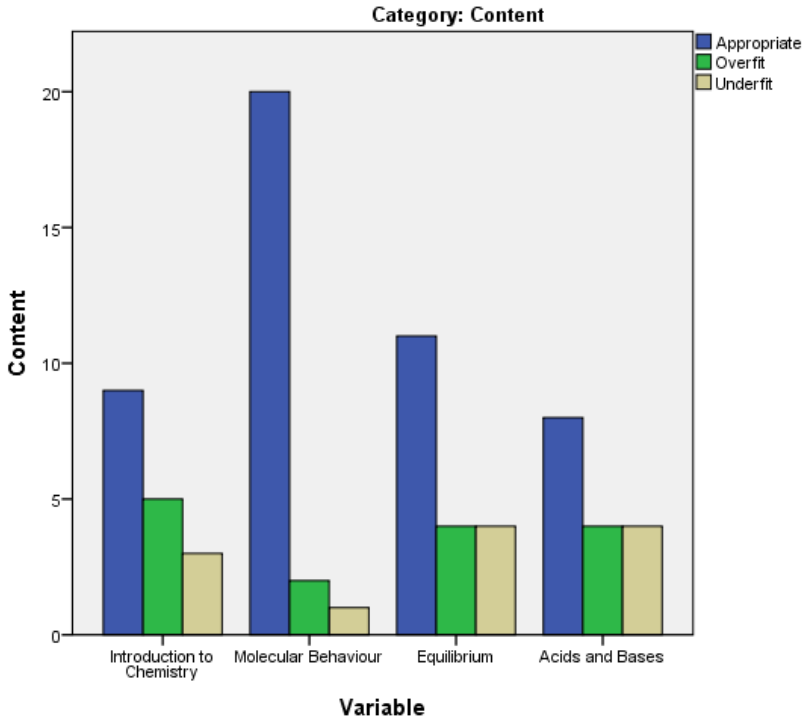

Figure 507: The Number of Unique Items Present for Each Topic Covered within Foundations of Chemistry IA from 2012-2015

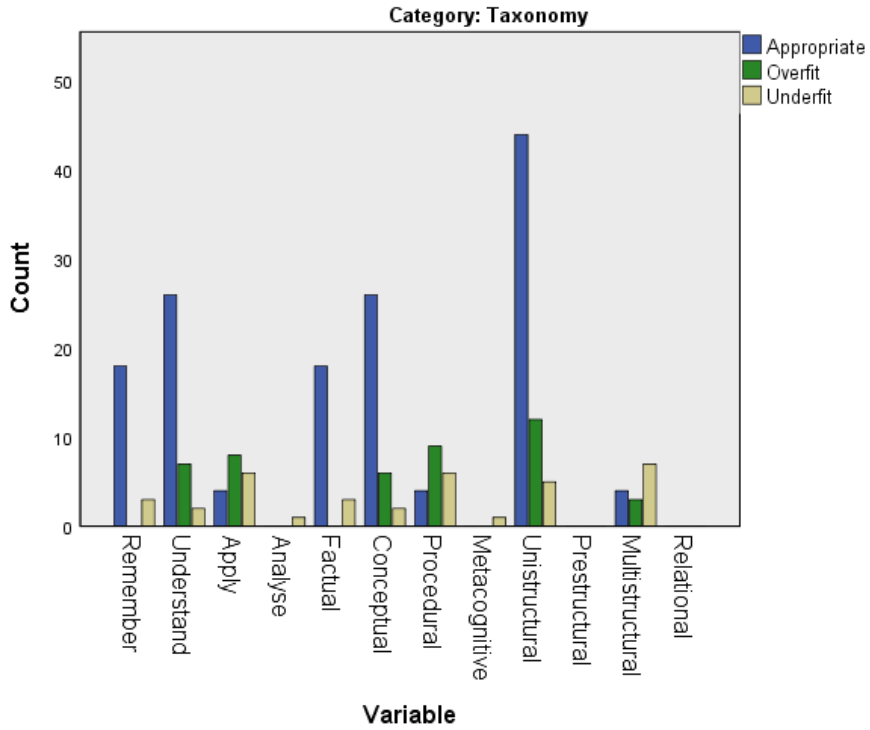

Figure 508: The Number of Unique Items Present for Each Taxonomy within Foundations of Chemistry IA from 2012-2015

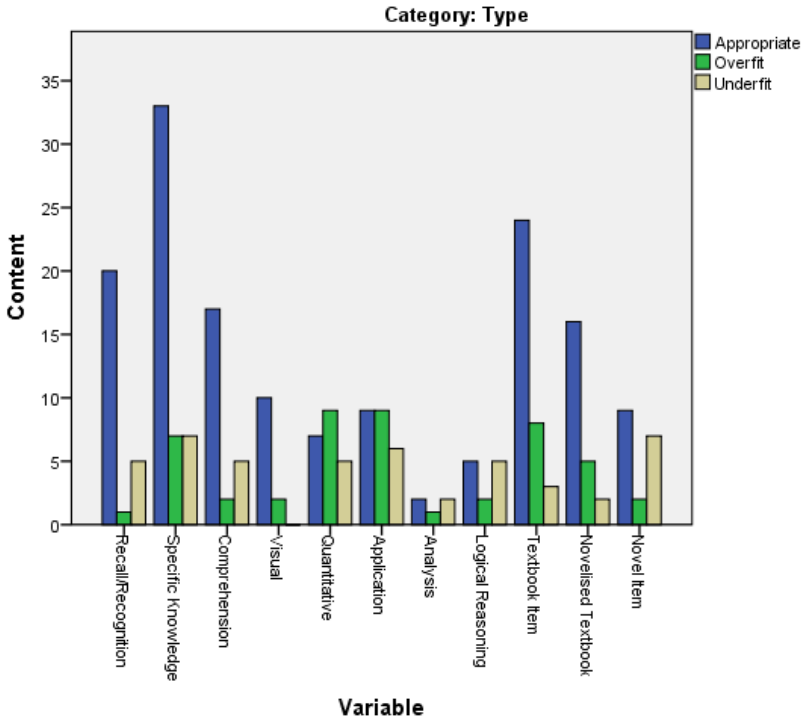

Figure 509: The Number of Unique Items Present for Each Item Type within Foundations of Chemistry IA from 2012-2015

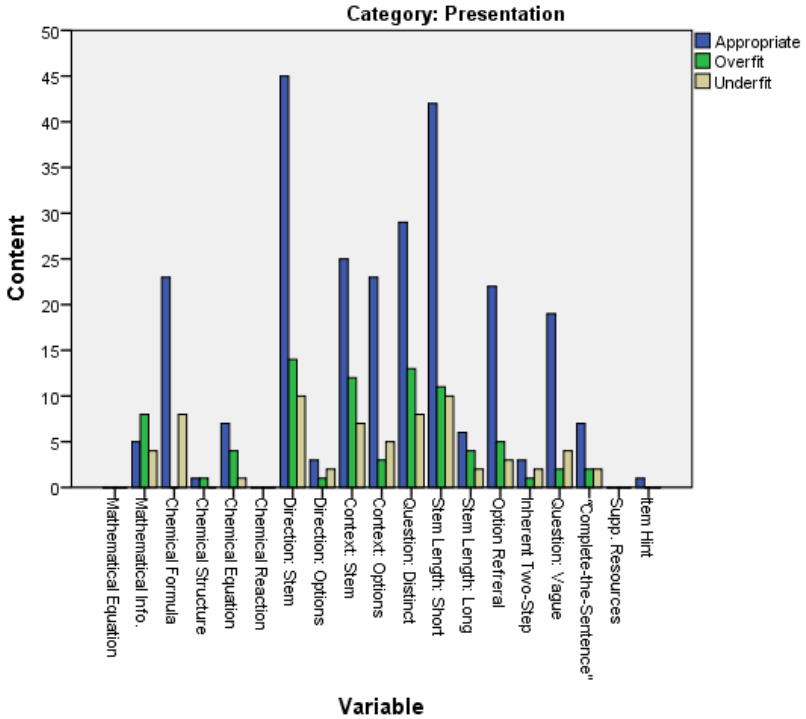

Figure 510: The Number of Unique Items Present for Each Item Presentation Style within Foundations of Chemistry IA from 2012-2015

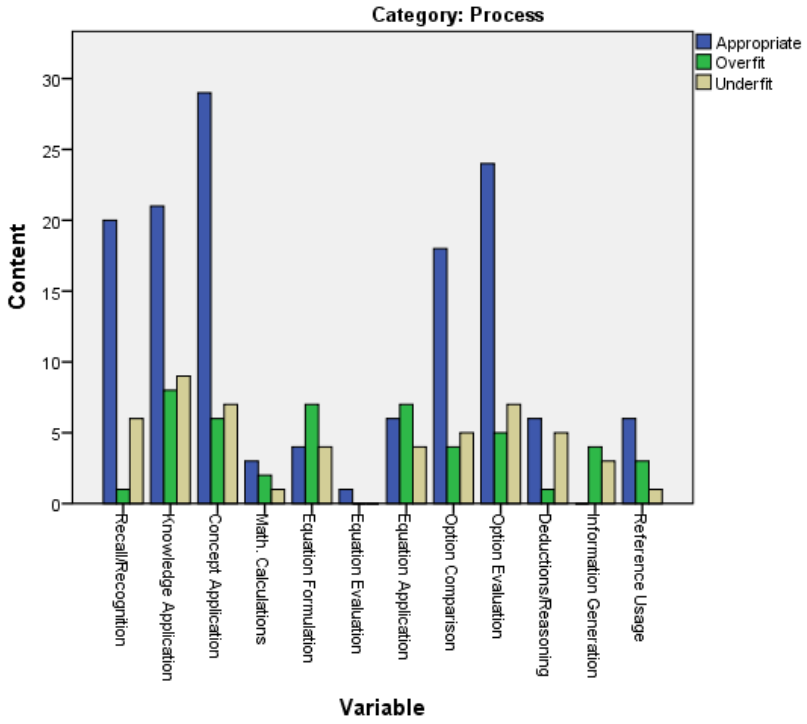

Figure 511: The Number of Unique Items Present for Each Item Process within Foundations of Chemistry IA from 2012-2015

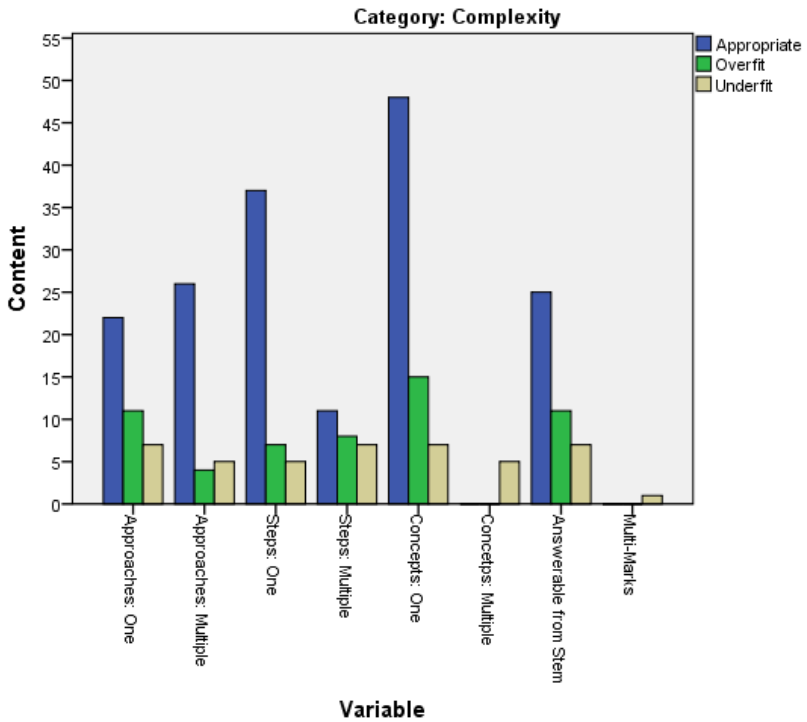

Figure 512: The Number of Unique Items Present for Each Level of Item Complexity within Foundations of Chemistry IA from 2012-2015

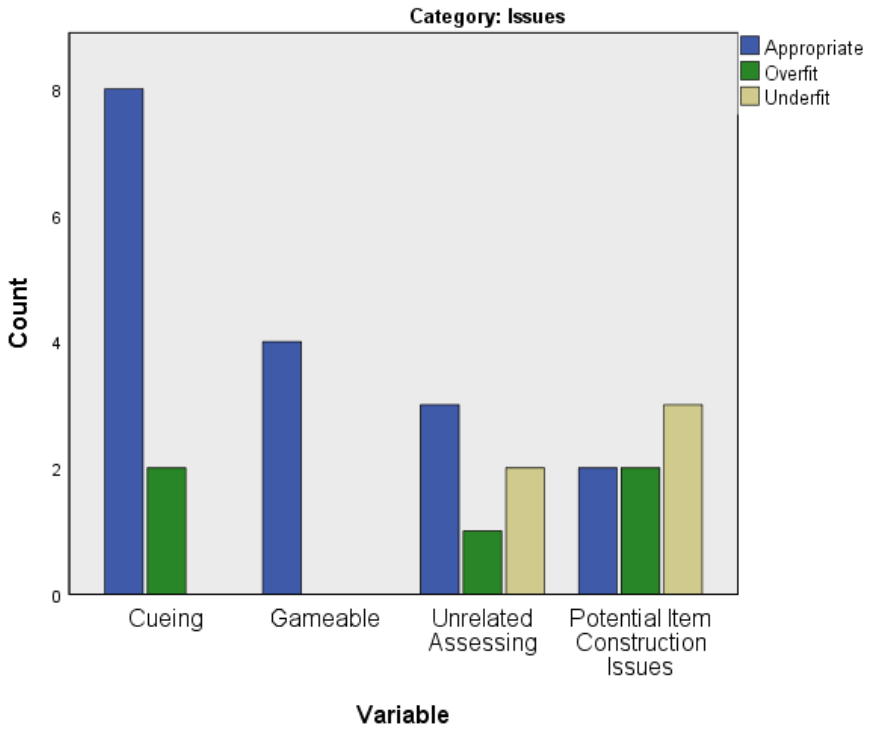

Figure 513: The Number of Unique Items Present with a Potential Item Flaws within Foundations of Chemistry IA from 2012-2015

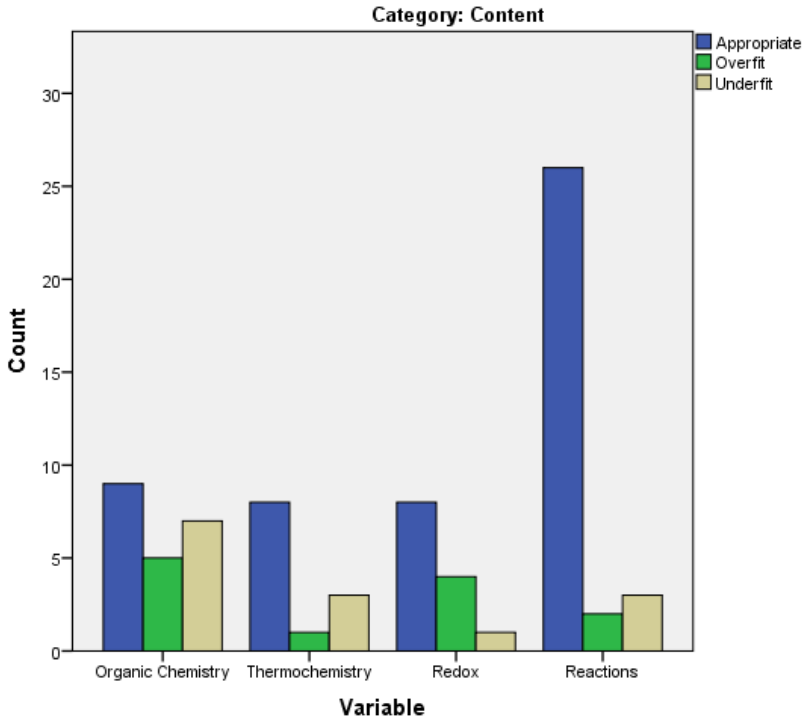

Figure 514: The Number of Unique Items Present for Each Topic Covered within Foundations of Chemistry IB from 2012-2015

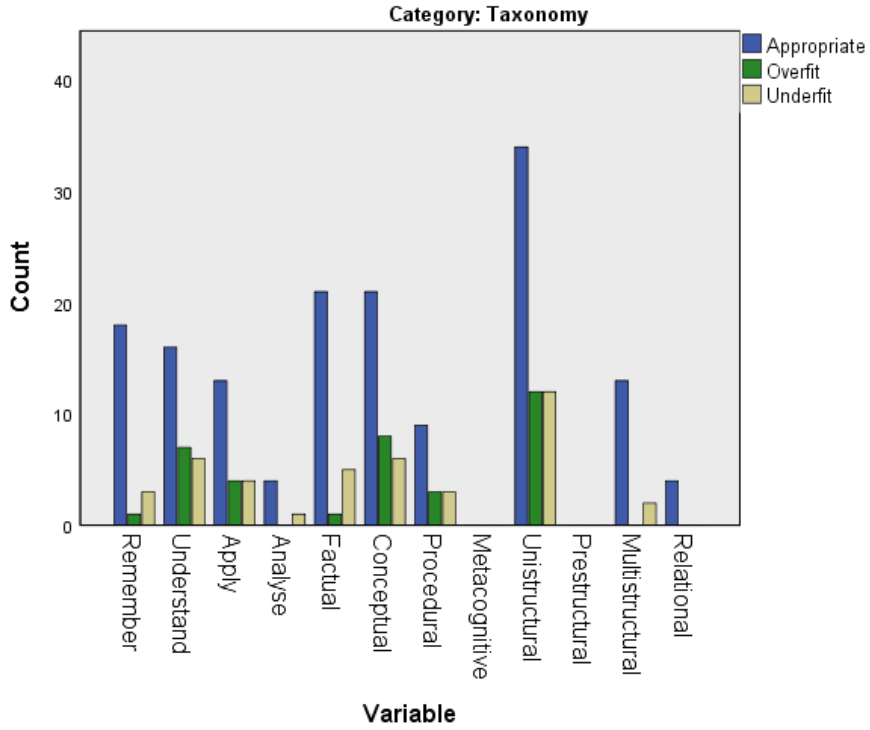

Figure 515: The Number of Unique Items Present for Each Taxonomy within Foundations of Chemistry IB from 2012-2015

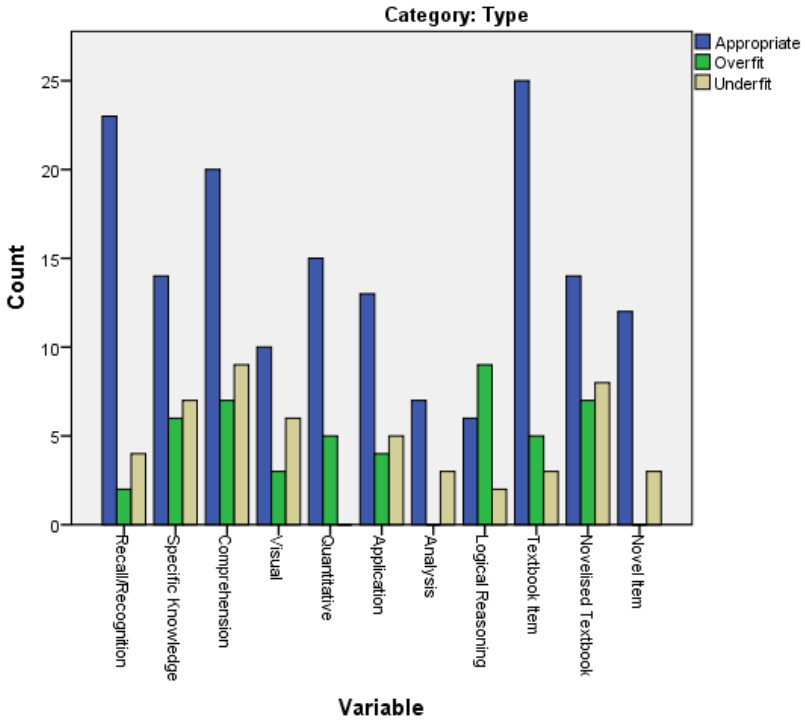

Figure 516: The Number of Unique Items Present for Each Item Type within Foundations of Chemistry IB from 2012-2015

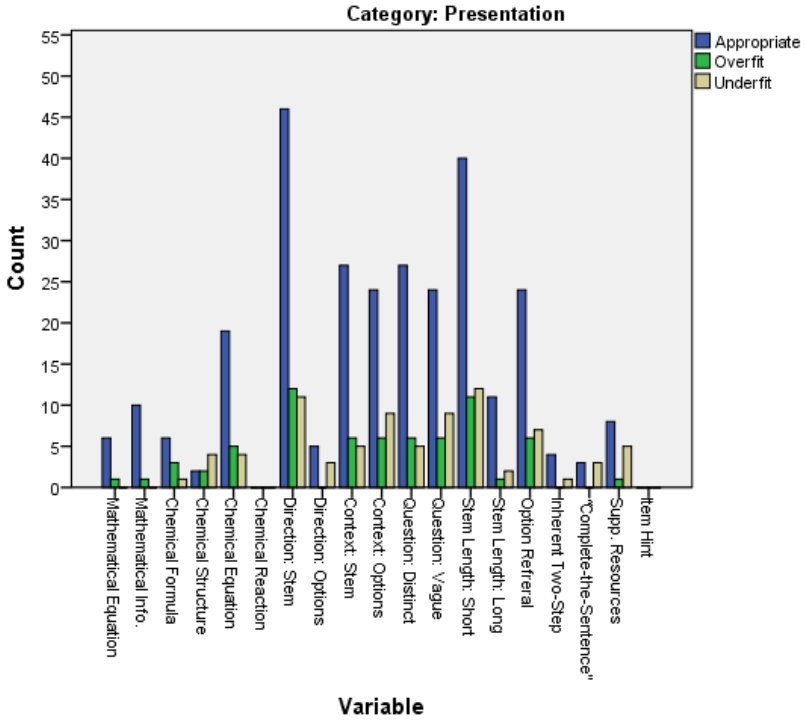

Figure 517: The Number of Unique Items Present for Each Item Presentation Style within Foundations of Chemistry IB from 2012-2015

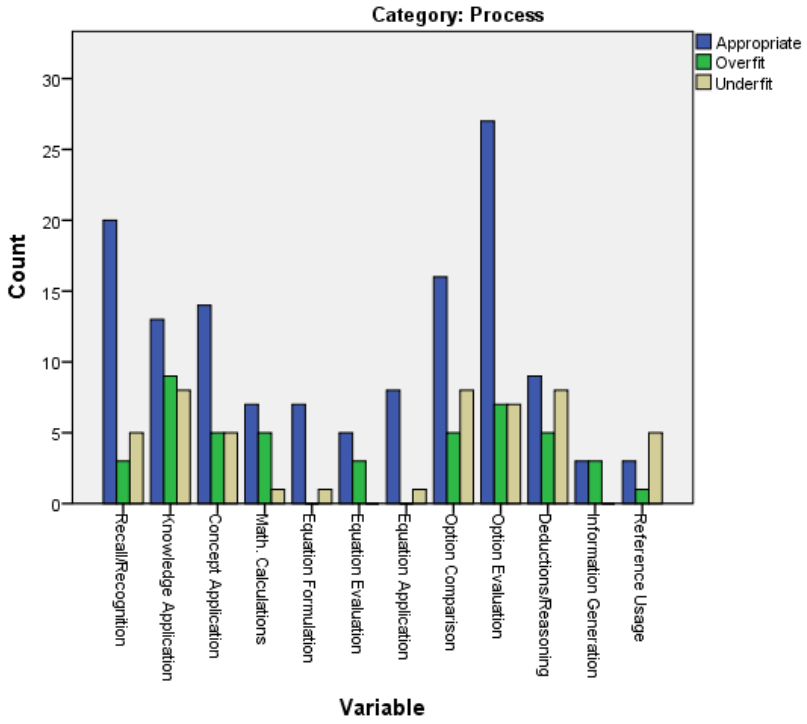

Figure 518: The Number of Unique Items Present for Each Item Process within Foundations of Chemistry IB from 2012-2015

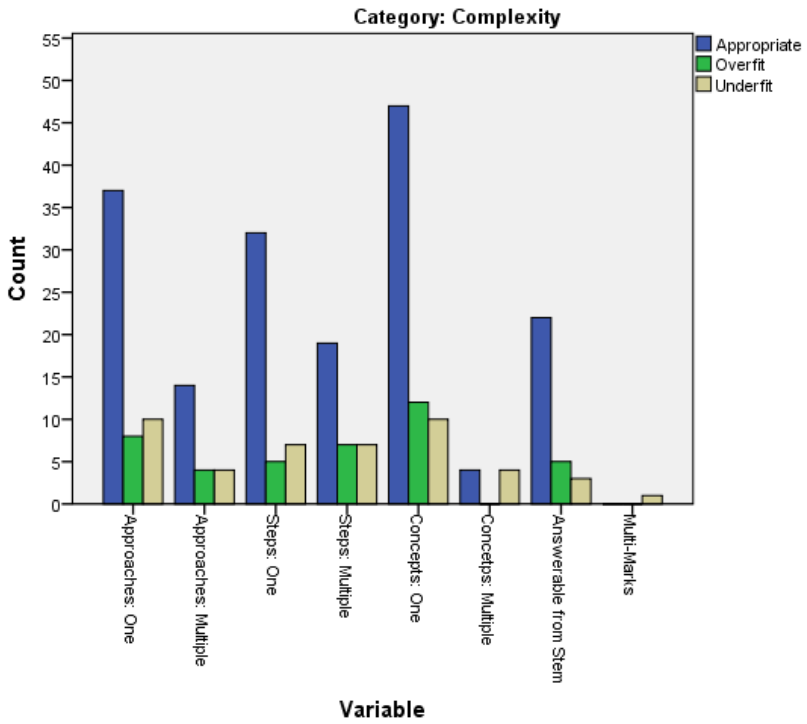

Figure 519: The Number of Unique Items Present for Each Item Complexity Level within Foundations of Chemistry IB from 2012-2015

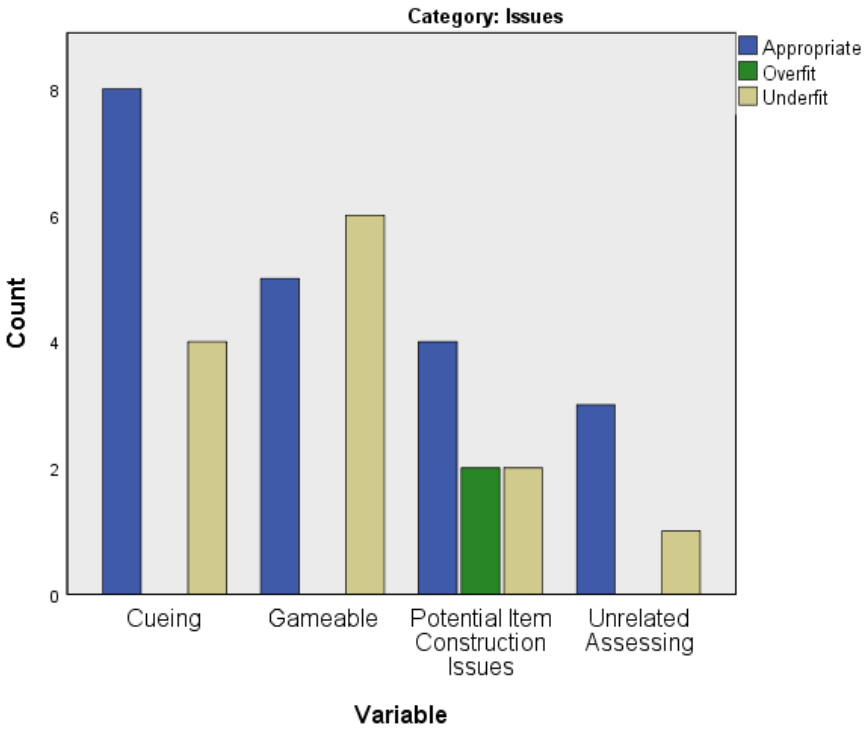

Figure 520: The Number of Unique Items Present with a Potential Item Flaws within Foundations of Chemistry IB from 2012-2015

7.17 Scatterplot Comparison of Student Raw Scores in Test-Retest MCQ Assessments

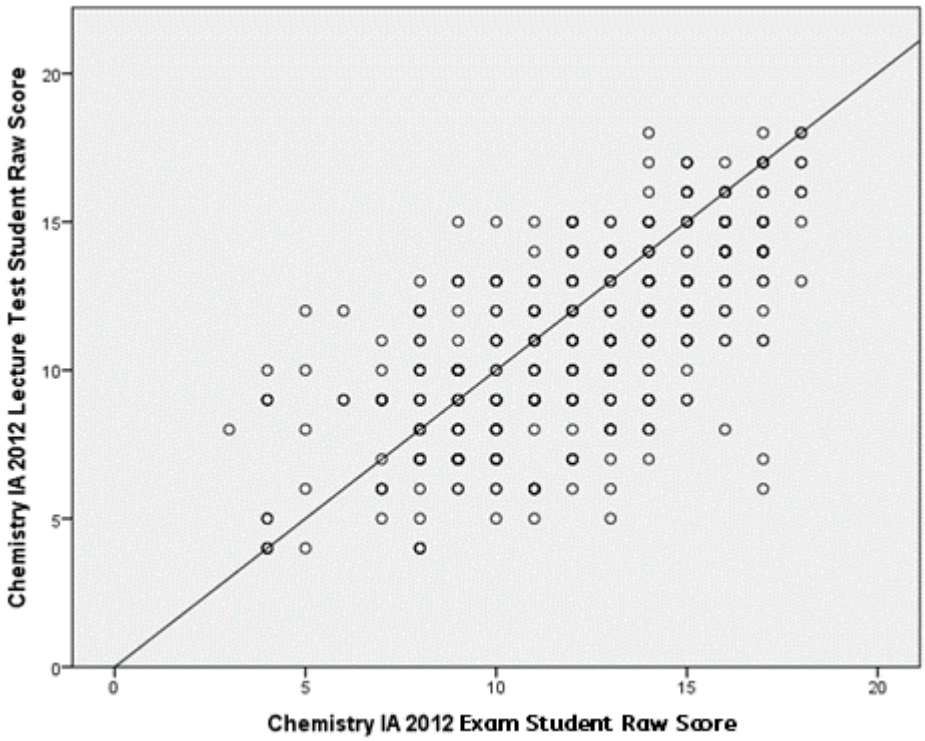

Figure 521: Scatterplot Comparison of Student Raw Scores Obtained Using the Same Items in Two Assessments within Chemistry IA during 2012 to View Changes in Student Performance

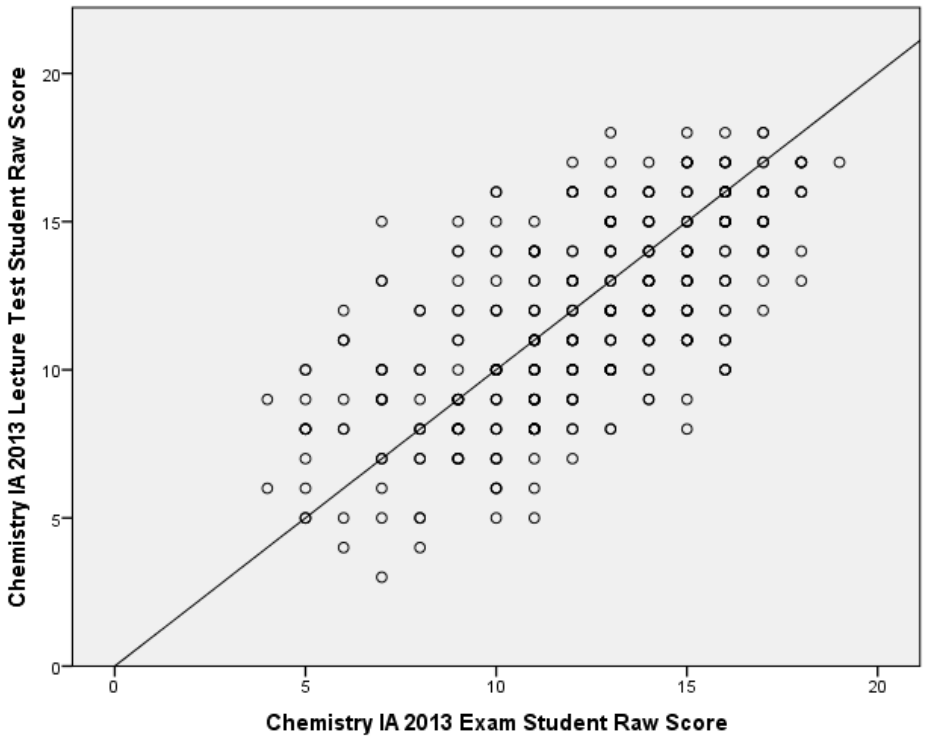

Figure 522: Scatterplot Comparison of Student Raw Scores Obtained Using the Same Items in Two Assessments within Chemistry IA during 2013 to View Changes in Student Performance

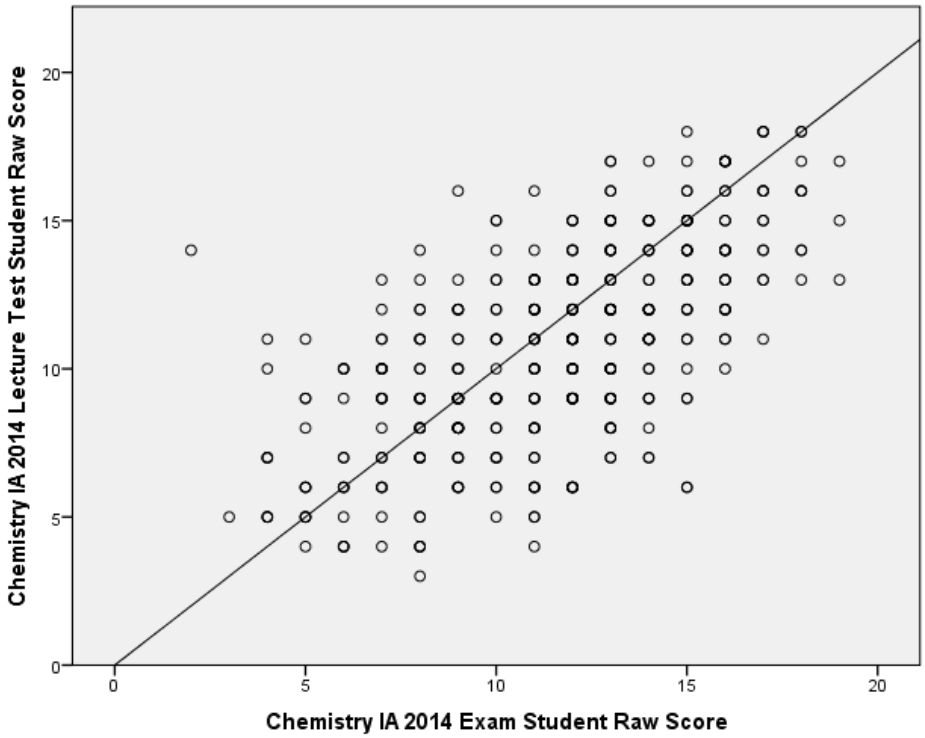

Figure 523: Scatterplot Comparison of Student Raw Scores Obtained Using the Same Items in Two Assessments within Chemistry IA during 2014 to View Changes in Student Performance

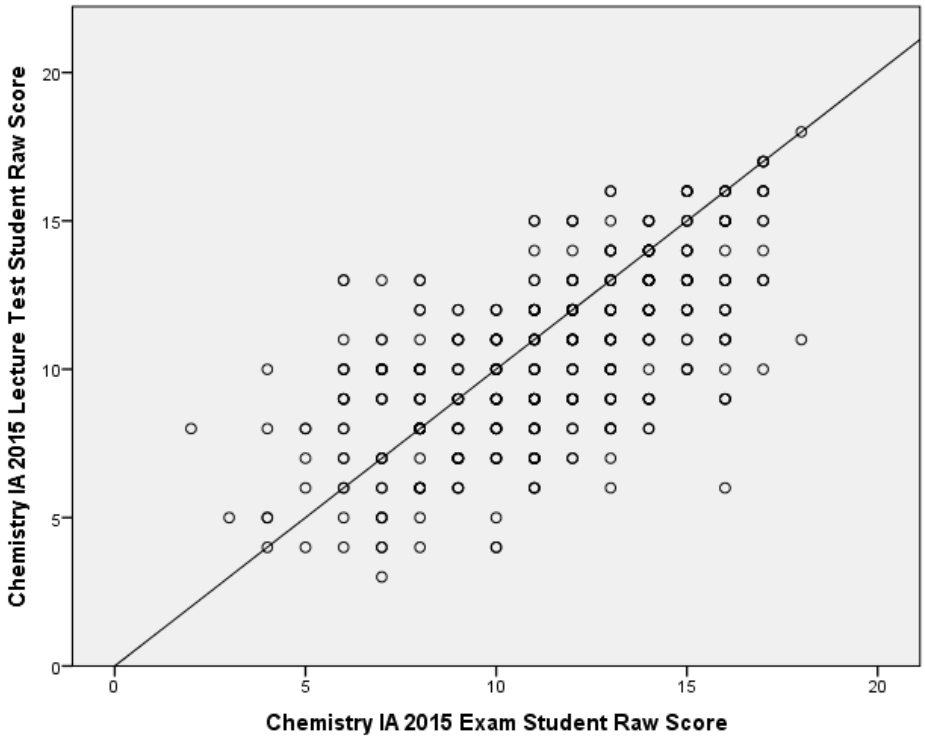

Figure 524: Scatterplot Comparison of Student Raw Scores Obtained Using the Same Items in Two Assessments within Chemistry IA during 2015 to View Changes in Student Performance

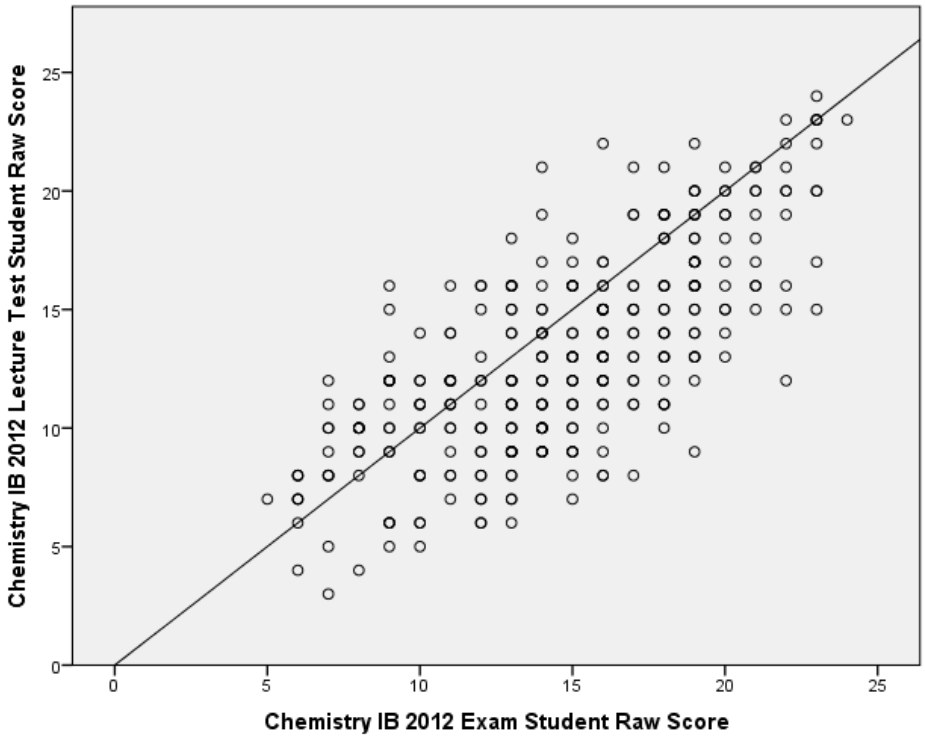

Figure 525: Scatterplot Comparison of Student Raw Scores Obtained Using the Same Items in Two Assessments within Chemistry IB during 2012 to View Changes in Student Performance

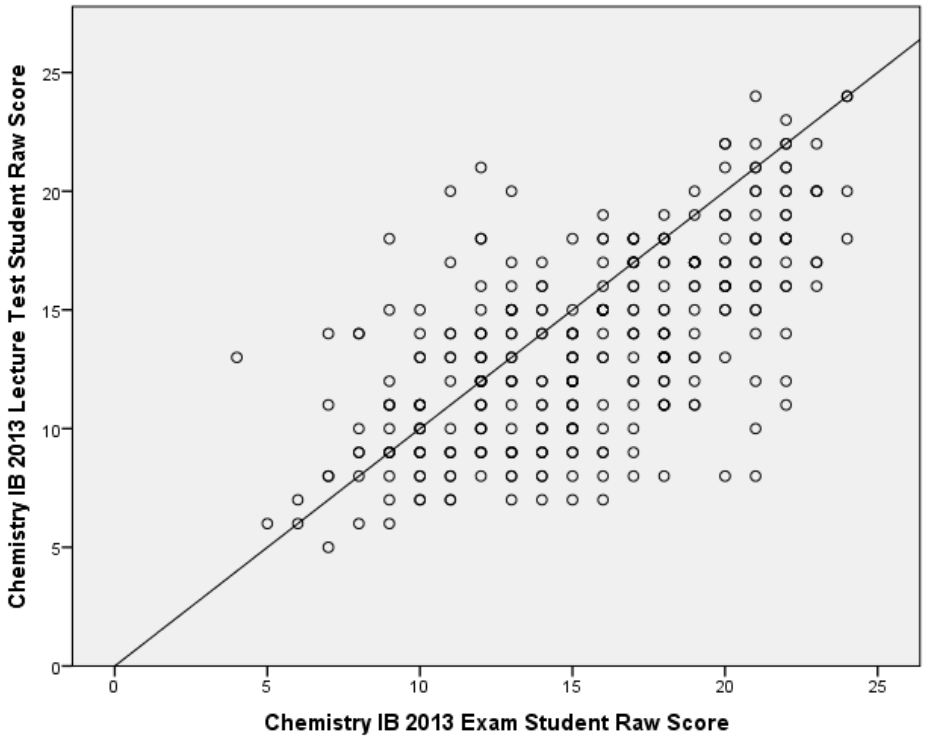

Figure 526: Scatterplot Comparison of Student Raw Scores Obtained Using the Same Items in Two Assessments within Chemistry IB during 2013 to View Changes in Student Performance

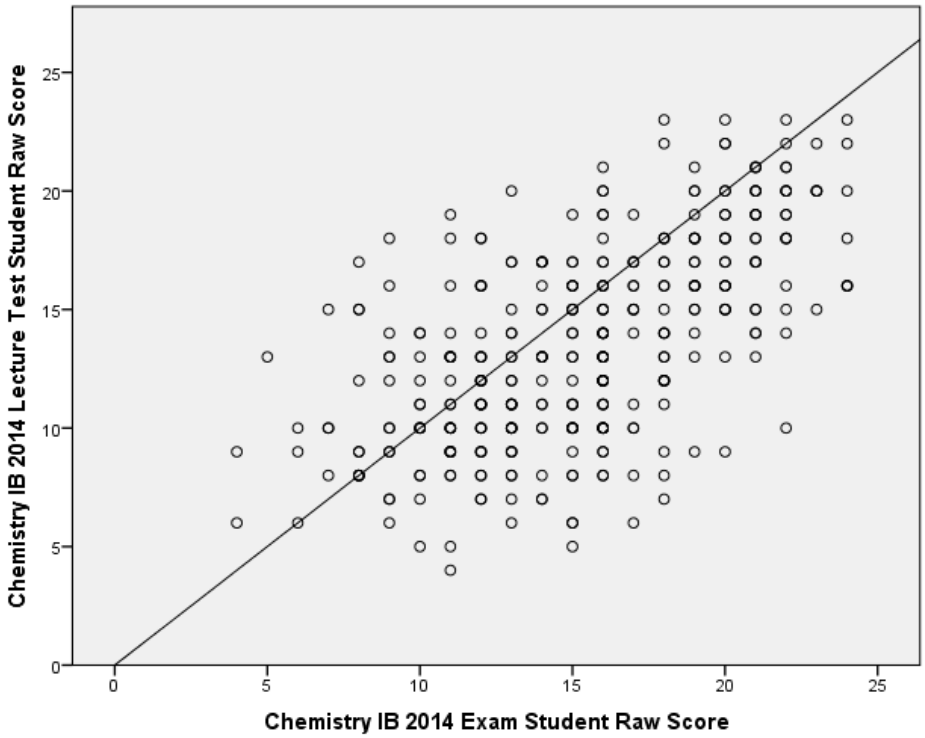

Figure 527: Scatterplot Comparison of Student Raw Scores Obtained Using the Same Items in Two Assessments within Chemistry IB during 2014 to View Changes in Student Performance

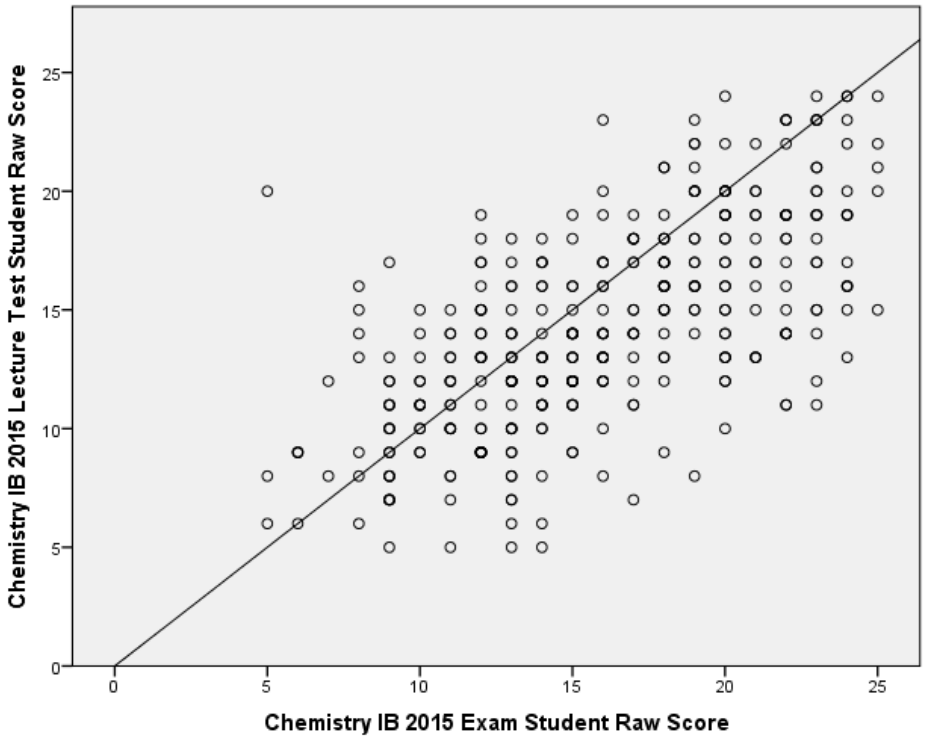

Figure 528: Scatterplot Comparison of Student Raw Scores Obtained Using the Same Items in Two Assessments within Chemistry IB during 2015 to View Changes in Student Performance

7.18 Distribution Comparison of Student Raw Scores in Test-Retest MCQ Assessments

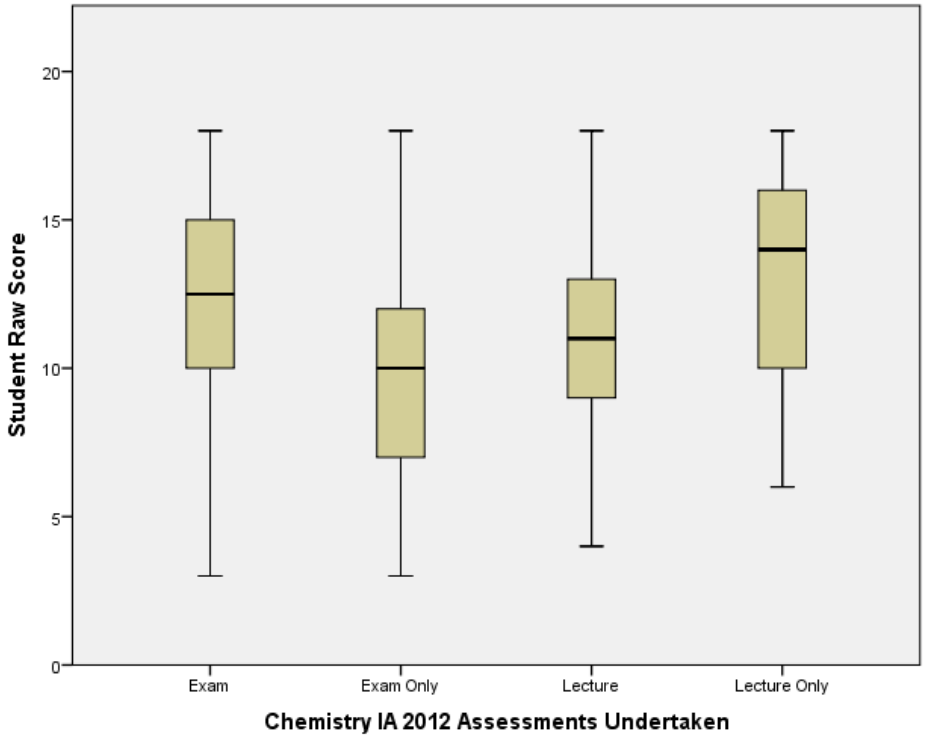

Figure 529: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IA During 2012, Separating Students Who Only Undertook One Assessment

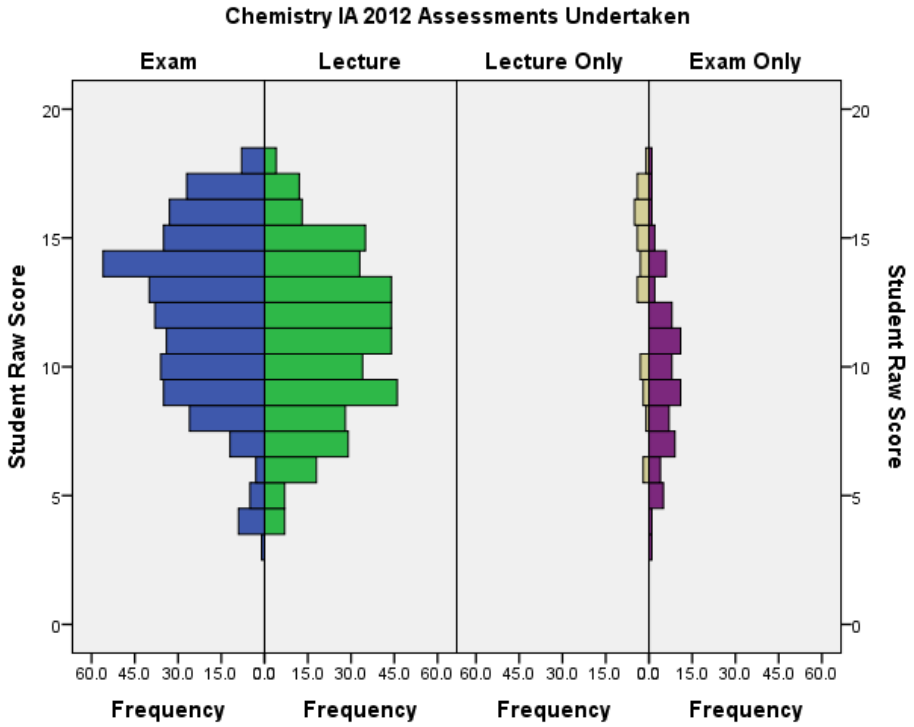

Figure 530: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA During 2012, Separating Student Who Only Undertook One Assessment

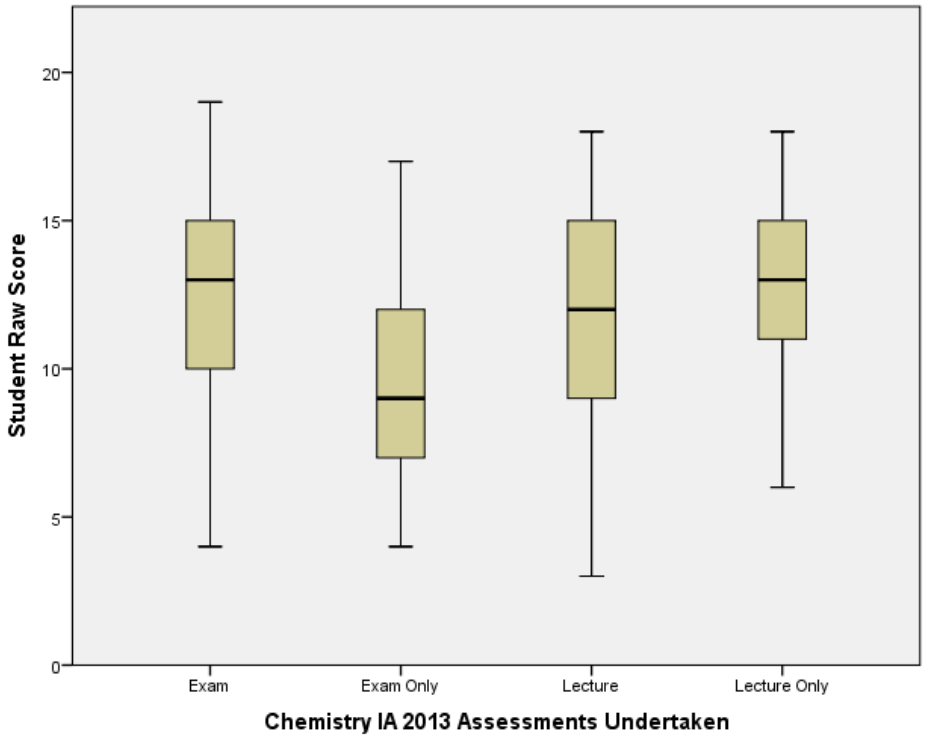

Figure 531: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IA During 2013, Separating Students Who Only Undertook One Assessment

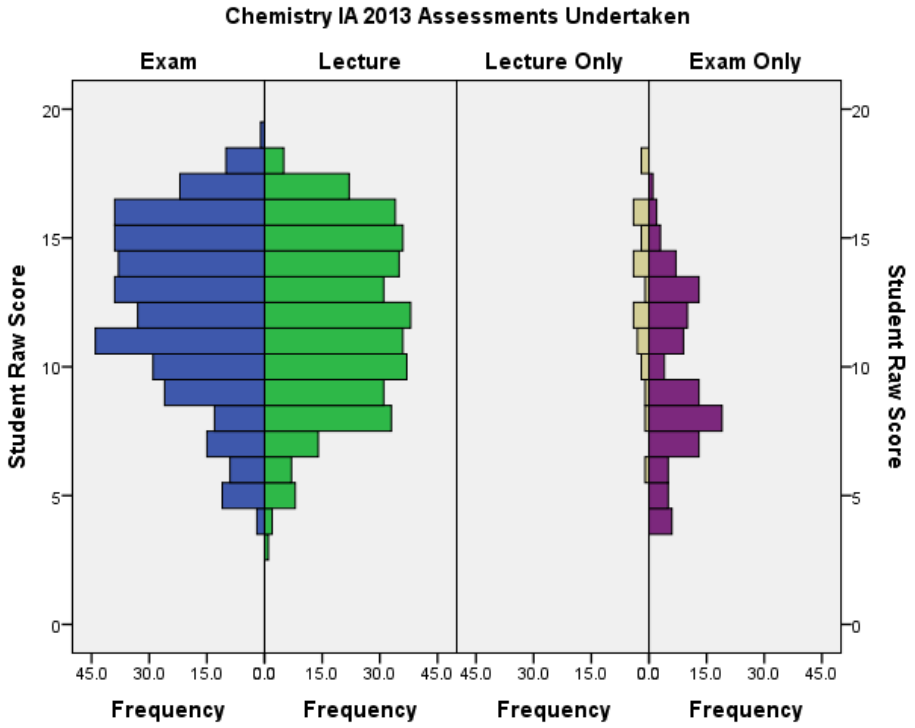

Figure 532: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA During 2013, Separating Student Who Only Undertook One Assessment

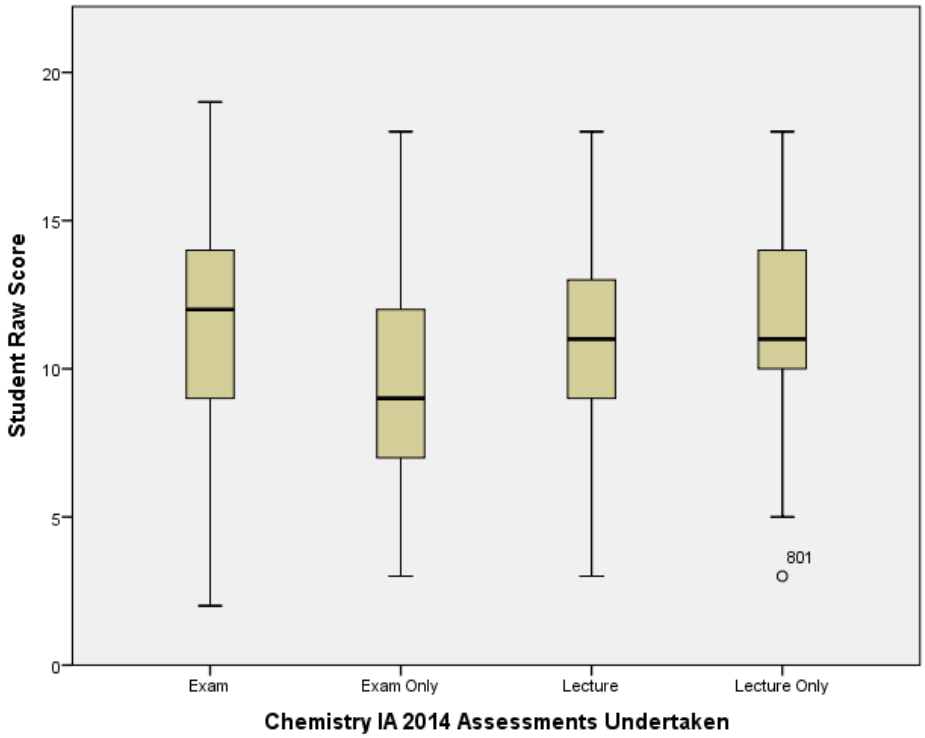

Figure 533: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IA During 2014, Separating Students Who Only Undertook One Assessment

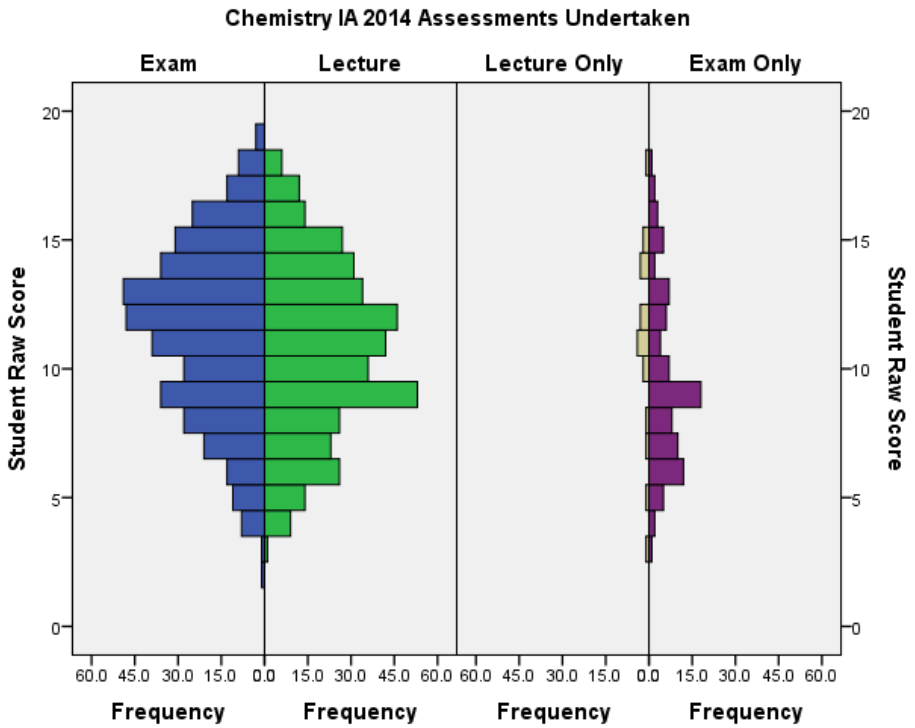

Figure 534: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA During 2014, Separating Student Who Only Undertook One Assessment

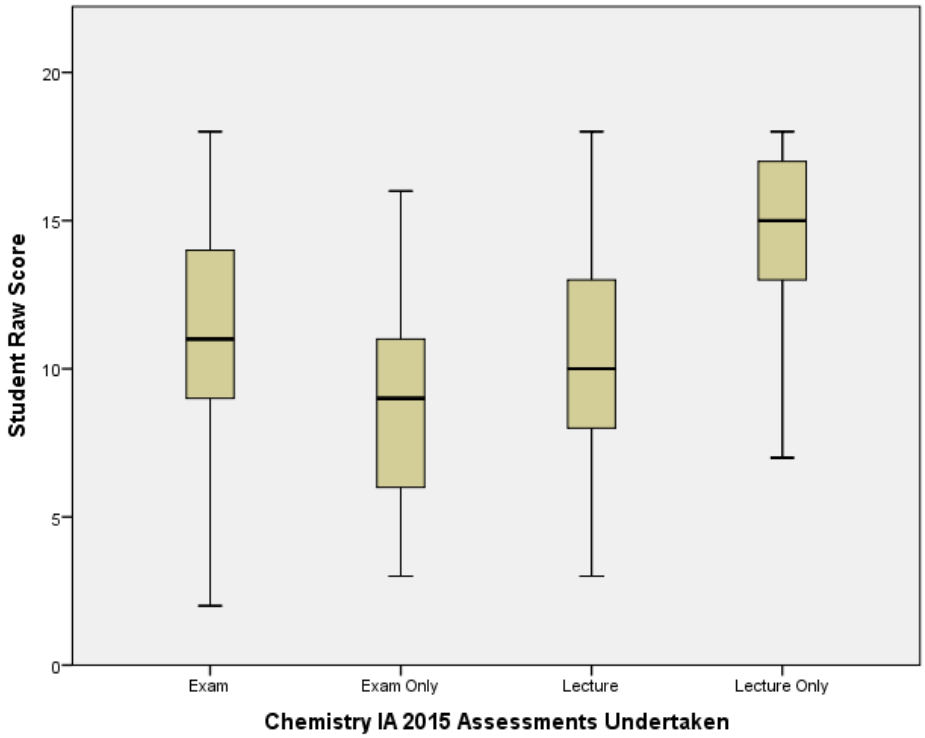

Figure 535: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IA During 2015, Separating Students Who Only Undertook One Assessment

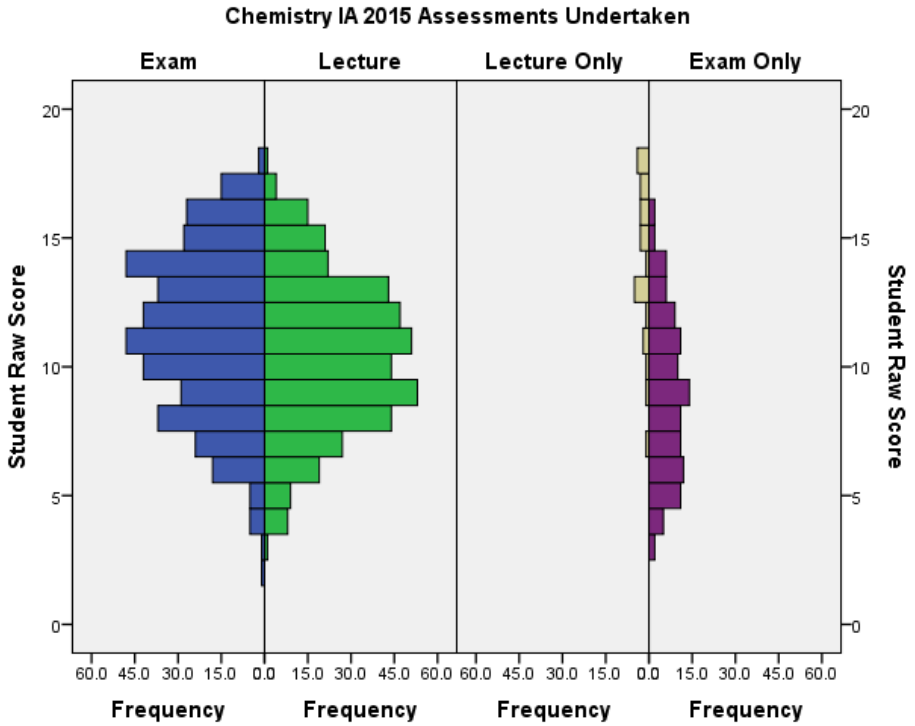

Figure 536: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA During 2015, Separating Student Who Only Undertook One Assessment

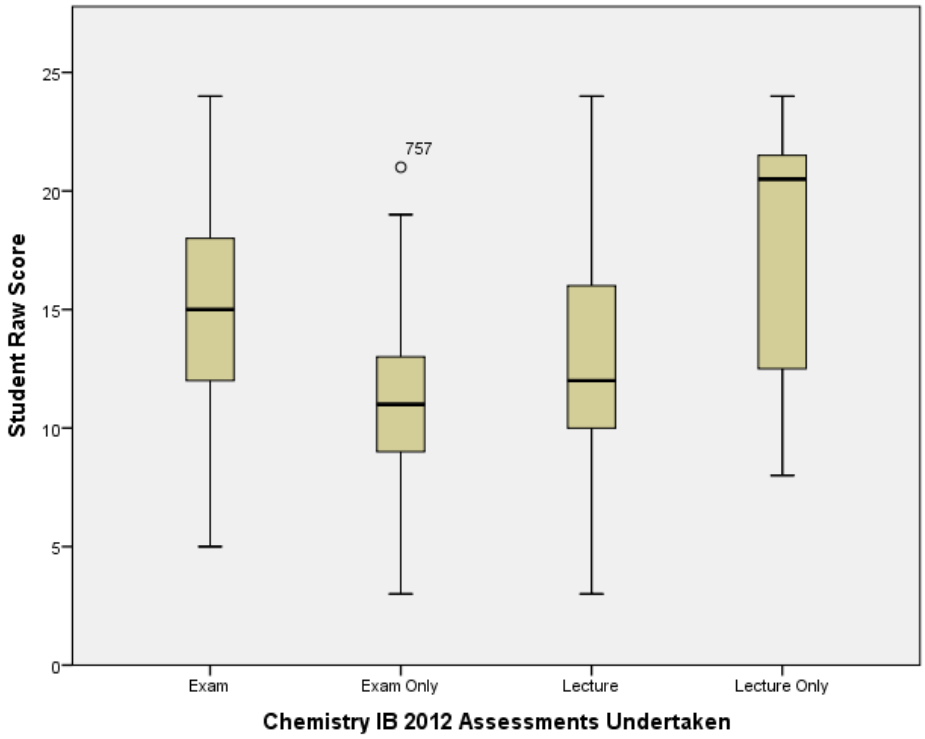

Figure 537: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IB During 2012, Separating Students Who Only Undertook One Assessment

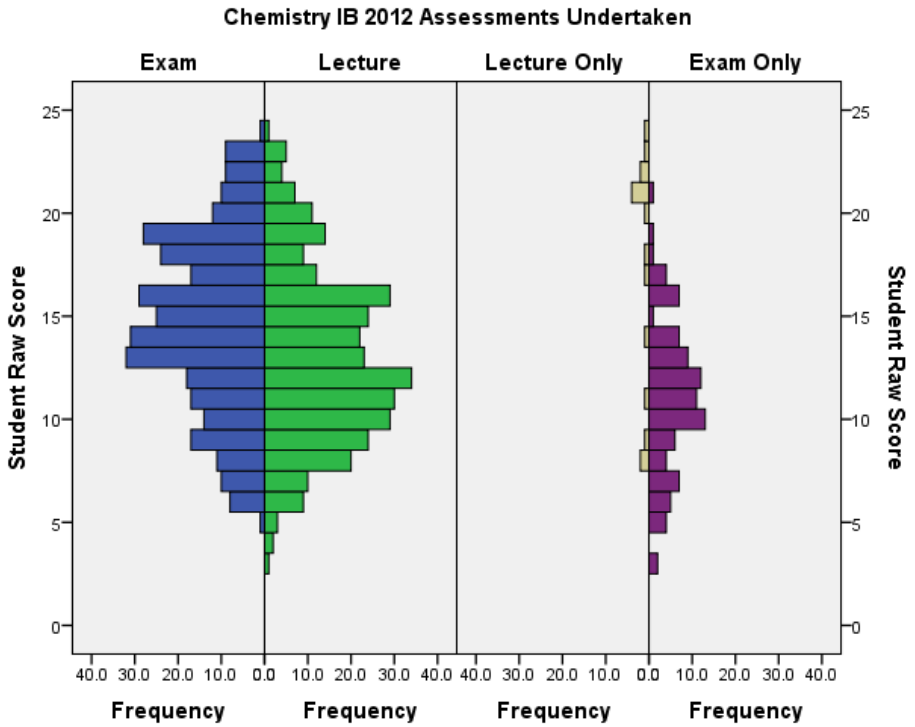

Figure 538: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IB During 2012, Separating Student Who Only Undertook One Assessment

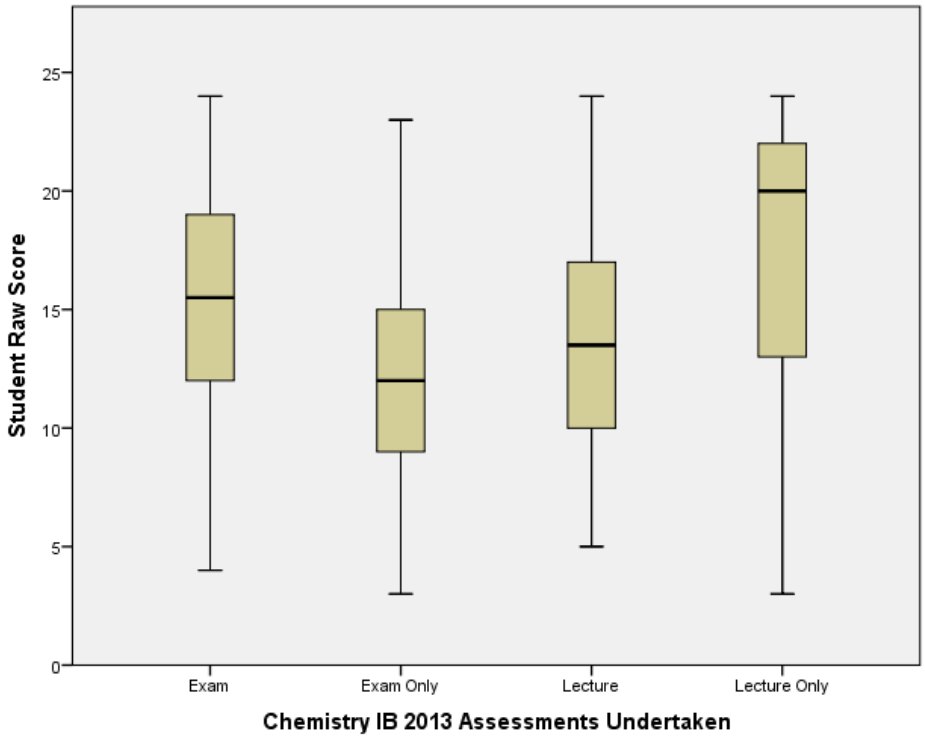

Figure 539: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IB During 2013, Separating Students Who Only Undertook One Assessment

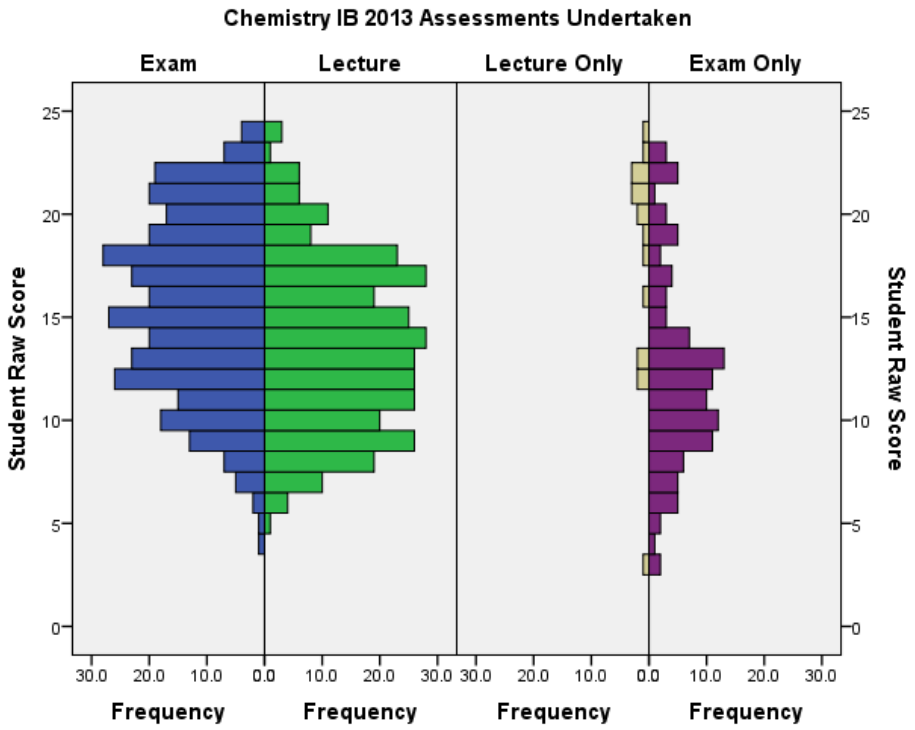

Figure 540: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IB During 2013, Separating Student Who Only Undertook One Assessment

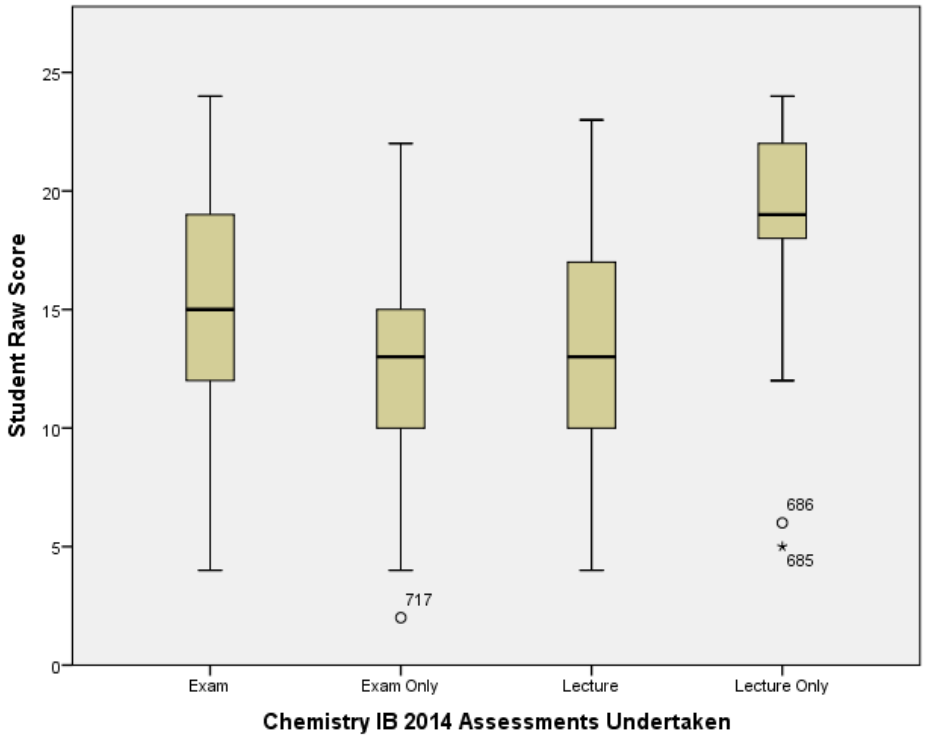

Figure 541: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IB During 2014, Separating Students Who Only Undertook One Assessment

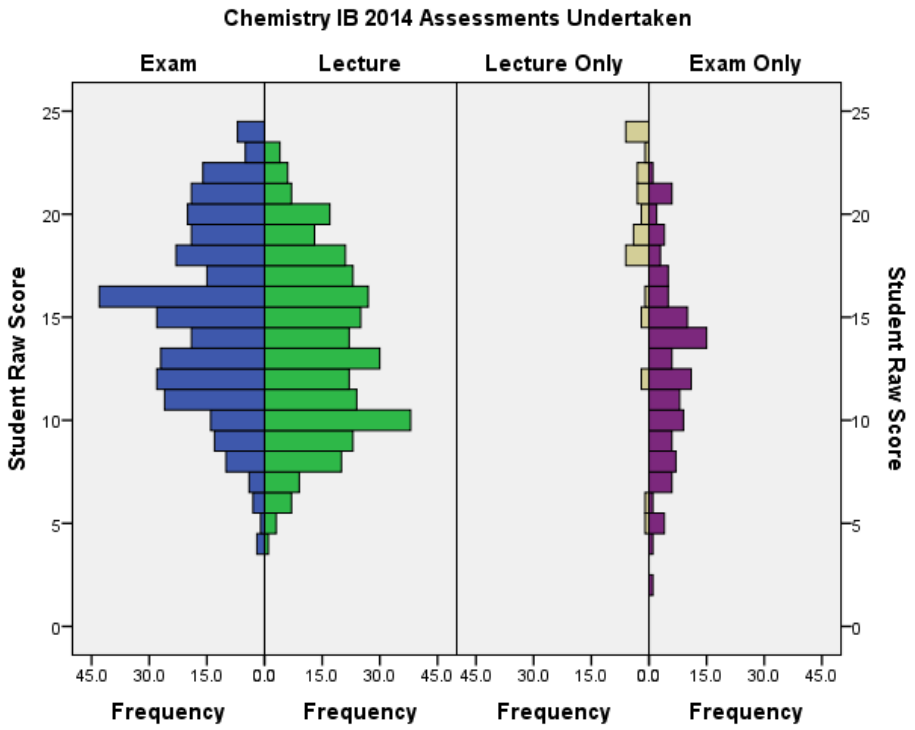

Figure 542: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IB During 2014, Separating Student Who Only Undertook One Assessment

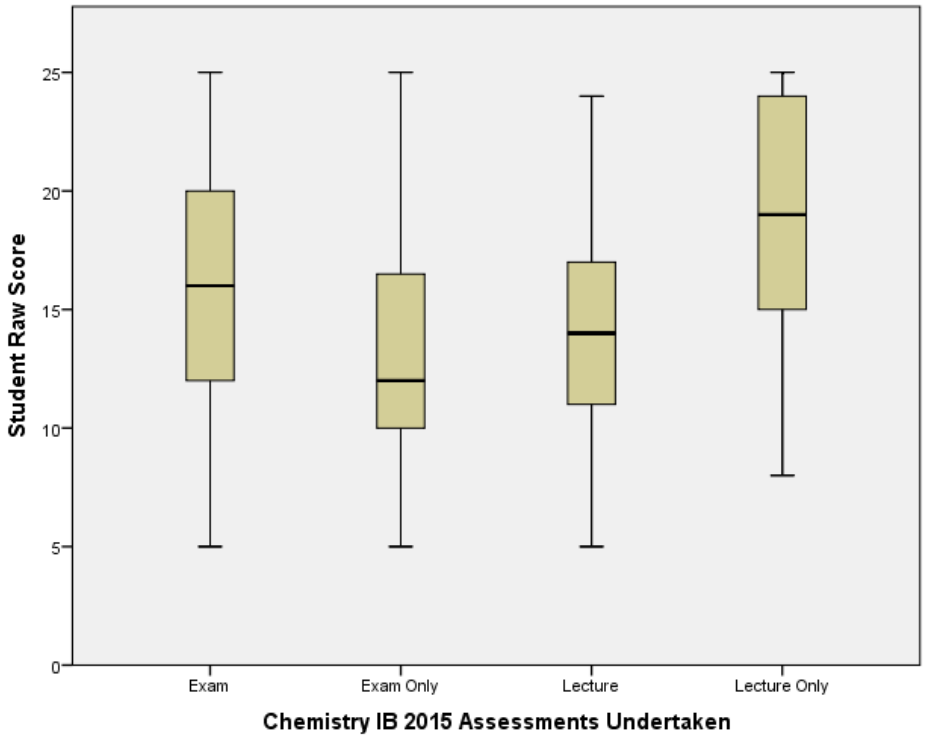

Figure 543: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IB During 2015, Separating Students Who Only Undertook One Assessment

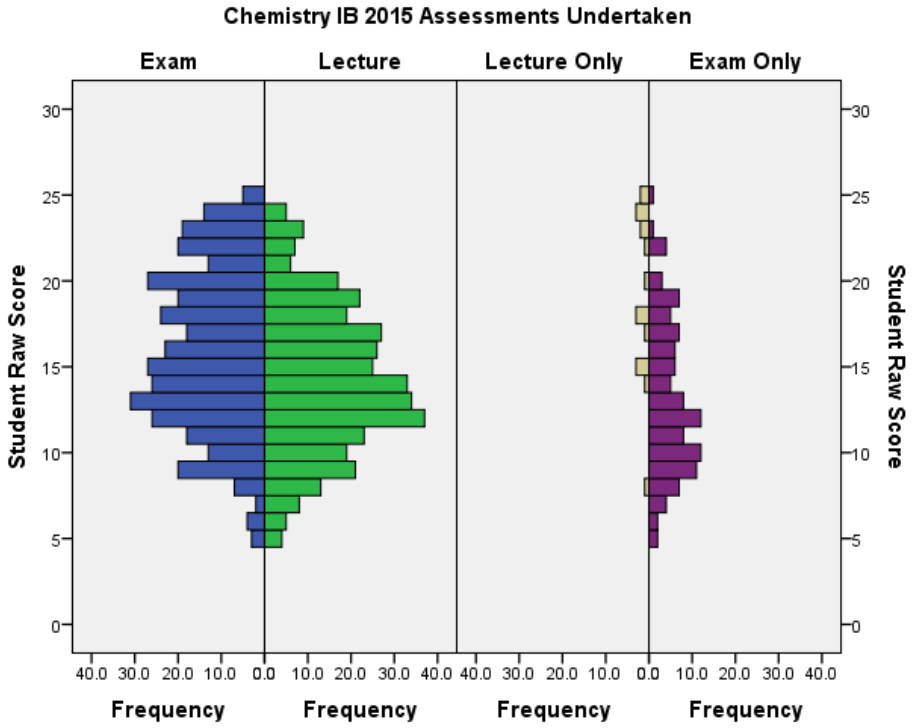

Figure 544: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IB During 2015, Separating Student Who Only Undertook One Assessment

7.19 Comparison of Changes in Student Raw Score Performance in Shared Items

Table 85: Raw Score Average Result from Overlapping Items within Chemistry IA MCQ Assessments from 2012 Separating Students Based on Their Shift in Performance

19 Item Overlap		Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	Cohort	506	11.07	3.17	12.16	3.28	
	Lecture Only	29	13.38	3.38			
	Exam Only	79			9.65	3.26	
Test-Retest Changes	Increase	247	10.32	2.98	13.04	2.78	2.73
	Decrease	105	12.33	2.76	10.03	3.16	-2.30
	No Change	46	12.26	0.08	12.26	0.08	

Table 86: Raw Score Average Result from Overlapping Items within Chemistry IA MCQ Assessments from 2013 Separating Students Based on Their Shift in Performance

19 Item Overlap		Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	Cohort	505	11.88	3.27	12.30	3.31	
	Lecture Only	25	12.92	2.98			
	Exam Only	110			9.63	3.14	
Test-Retest Changes	Increase	198	10.97	3.11	13.34	2.87	2.37
	Decrease	118	13.41	3.00	10.73	3.42	-2.68
	No Change	54	11.91	0.09	11.91	0.09	

Table 87: Raw Score Average Result from Overlapping Items within Chemistry IA MCQ Assessments from 2014 Separating Students Based on Their Shift in Performance

19 Item Overlap		Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	Cohort	512	10.76	3.34	11.49	3.43	
	Lecture Only	19	11.21	3.53			
	Exam Only	93			9.51	3.40	
Test-Retest Changes	Increase	216	9.78	3.07	12.59	2.94	2.81
	Decrease	132	12.21	3.16	9.84	3.49	-2.37
	No Change	52	11.12	0.08	11.12	0.08	

Table 88: Raw Score Average Result from Overlapping Items within Chemistry IA MCQ Assessments from 2015 Separating Students Based on Their Shift in Performance

18 Item Overlap		Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	Cohort	546	10.48	2.95	11.32	3.22	
	Lecture Only	25	14.20	2.98			
	Exam Only	112			8.90	3.11	
Test-Retest Changes	Increase	222	9.74	2.85	12.43	2.76	2.69
	Decrease	113	11.33	2.56	9.11	2.94	-2.22
	No Change	74	11.39	0.10	11.39	0.10	

Table 89: Raw Score Average Result from Overlapping Items within Chemistry IB MCQ Assessments from 2012 Separating Students Based on Their Shift in Performance

24 Item Overlap		Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	Cohort	434	13.04	4.23	14.64	4.30	
	Lecture Only	16	17.50	5.45			
	Exam Only	95			11.09	3.61	
Test-Retest Changes	Increase	200	12.19	3.84	15.81	3.57	3.62
	Decrease	92	14.04	4.25	11.79	4.18	-2.25
	No Change	31	15.52	0.09	15.52	0.09	

Table 90: Raw Score Average Result from Overlapping Items within Chemistry IB MCQ Assessments from 2013 Separating Students Based on Their Shift in Performance

24 Item Overlap		Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	Cohort	448	13.63	4.05	15.51	4.36	
	Lecture Only	18	17.89	5.25			
	Exam Only	114			12.39	4.71	
Test-Retest Changes	Increase	205	12.90	3.80	16.80	3.76	3.90
	Decrease	76	14.96	3.86	12.24	3.92	-2.72
	No Change	35	15.03	0.09	15.03	0.09	

Table 91: Raw Score Average Result from Overlapping Items within Chemistry IB MCQ Assessments from 2014 Separating Students Based on Their Shift in Performance

24 Item Overlap		Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	Cohort	485	13.54	4.21	15.23	4.32	
	Lecture Only	32	18.78	4.73			
	Exam Only	111			12.75	4.32	
Test-Retest Changes	Increase	219	12.74	4.08	16.60	3.76	3.86
	Decrease	85	15.38	3.79	12.20	3.91	-3.18
	No Change	38	14.05	0.09	14.05	0.09	

Table 92: Raw Score Average Result from Overlapping Items within Chemistry IB MCQ Assessments from 2015 Separating Students Based on Their Shift in Performance

25 Item Overlap		Count	Lecture Test Average Score	Lecture Test S.D.	Exam Average Score	Exam S.D.	Average Score Change
Student Cohorts	Cohort	489	14.30	4.25	15.94	4.74	
	Lecture Only	18	19.33	4.64			
	Exam Only	111			13.11	4.41	
Test-Retest Changes	Increase	226	13.45	3.93	17.27	4.26	3.82
	Decrease	102	15.64	4.11	12.96	4.26	-2.68
	No Change	32	16.03	0.09	16.03	0.09	

7.20 Histogram Comparison of Student Raw Scores in Test-Retest Assessments

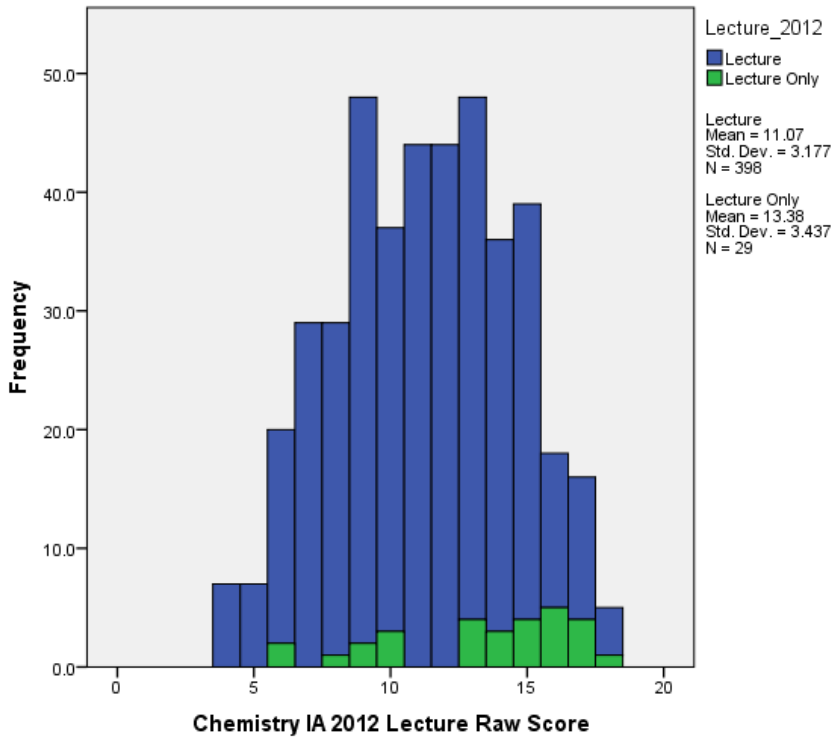

Figure 545: The Results of Students on Overlapping Items within Chemistry IA Lecture Test MCQ Assessments in 2012 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

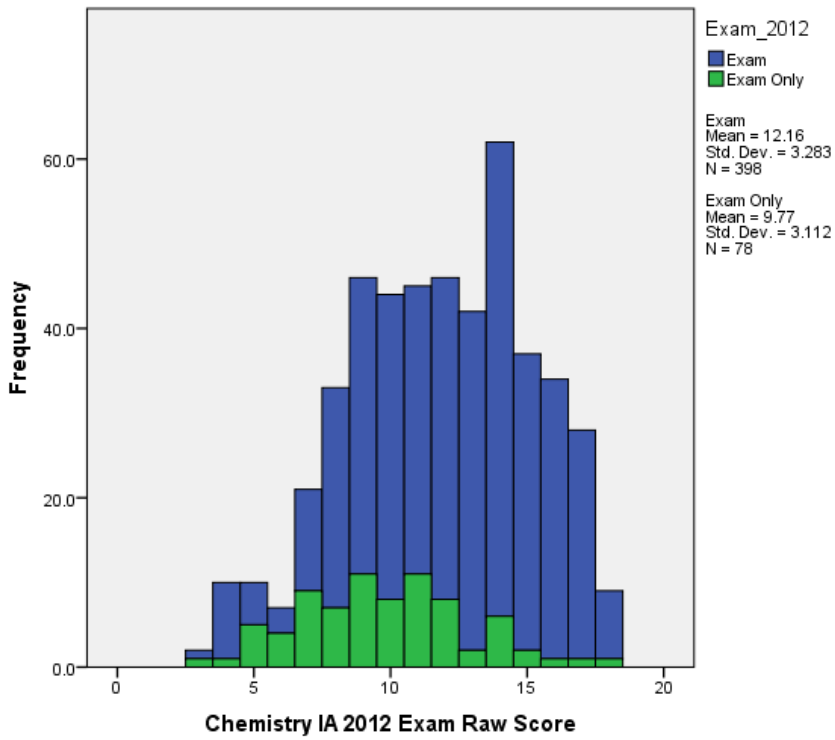

Figure 546: The Results of Students on Overlapping Items within Chemistry IA Redeemable Exam MCQ Assessment in 2012 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

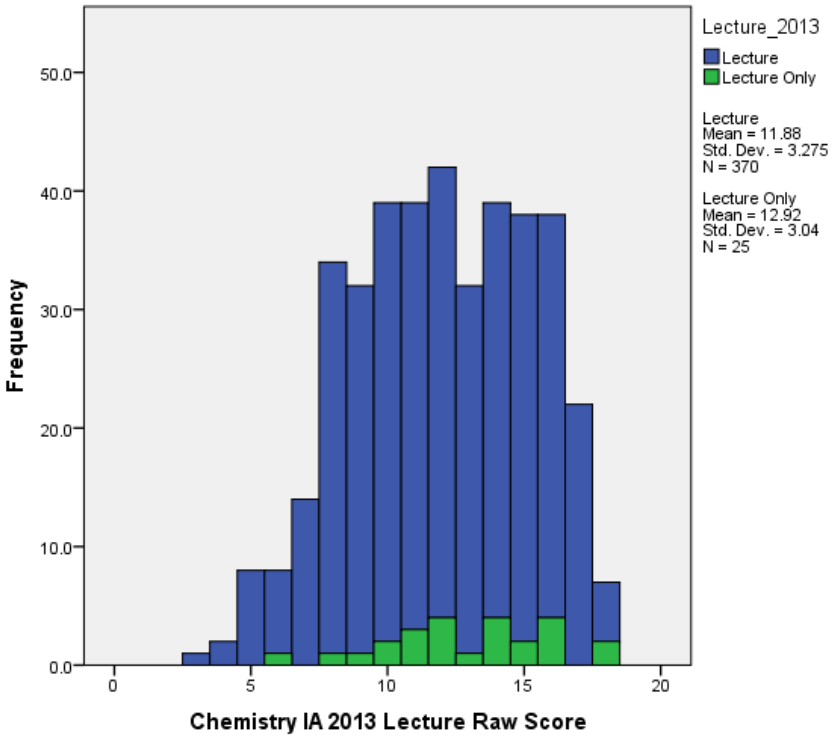

Figure 547: The Results of Students on Overlapping Items within Chemistry IA Lecture Test MCQ Assessments in 2013 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

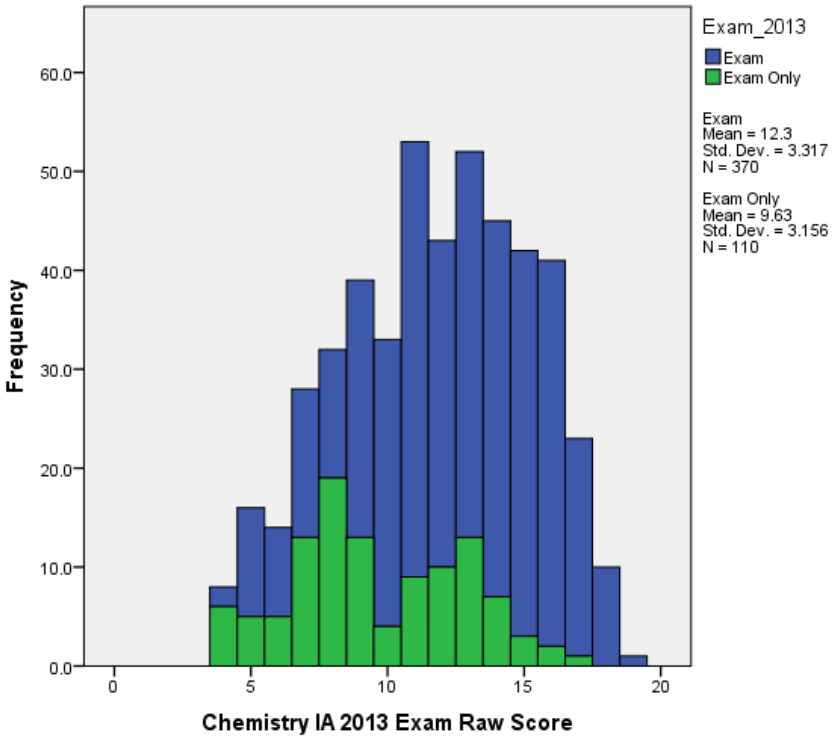

Figure 548: The Results of Students on Overlapping Items within Chemistry IA Redeemable Exam MCQ Assessment in 2013 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

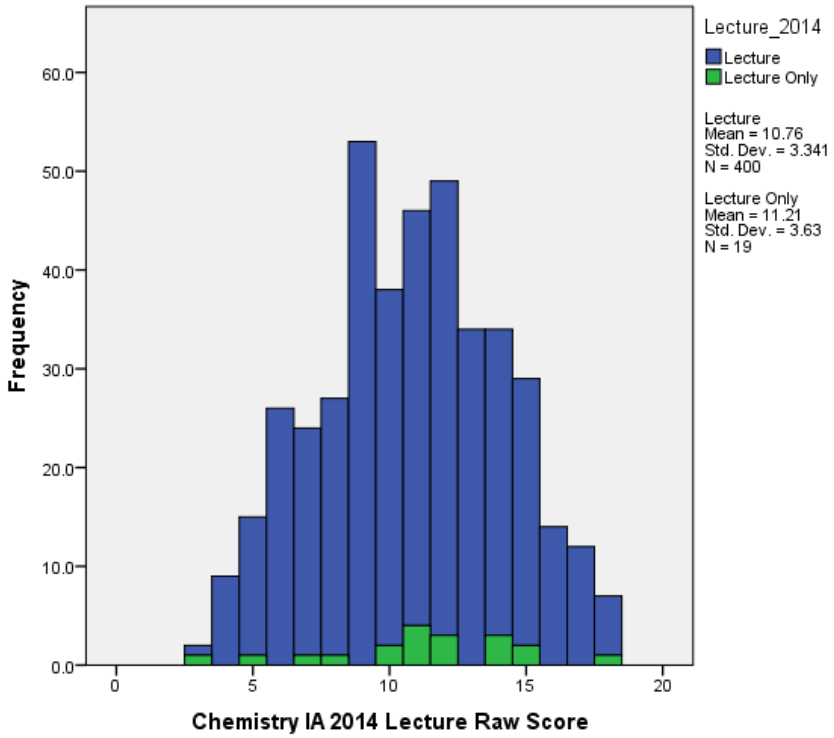

Figure 549: The Results of Students on Overlapping Items within Chemistry IA Lecture Test MCQ Assessments in 2014 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

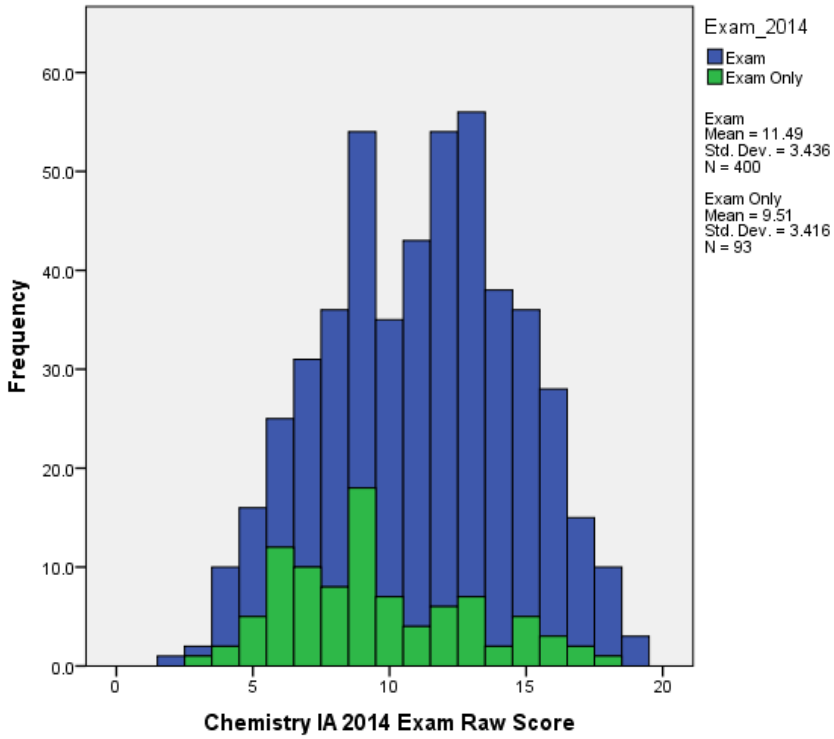

Figure 550: The Results of Students on Overlapping Items within Chemistry IA Redeemable Exam MCQ Assessment in 2014 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

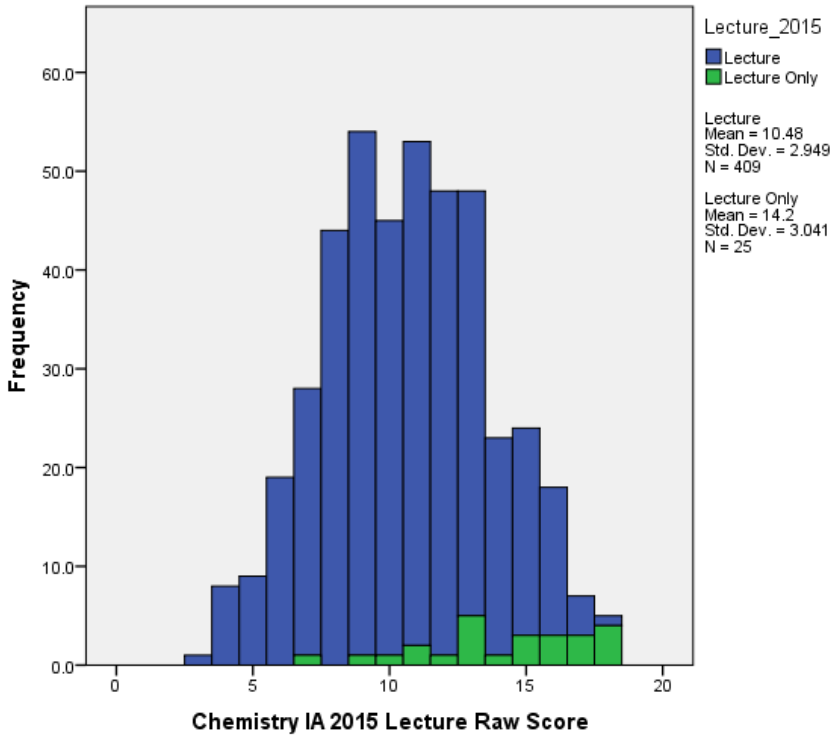

Figure 551: The Results of Students on Overlapping Items within Chemistry IA Lecture Test MCQ Assessments in 2015 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

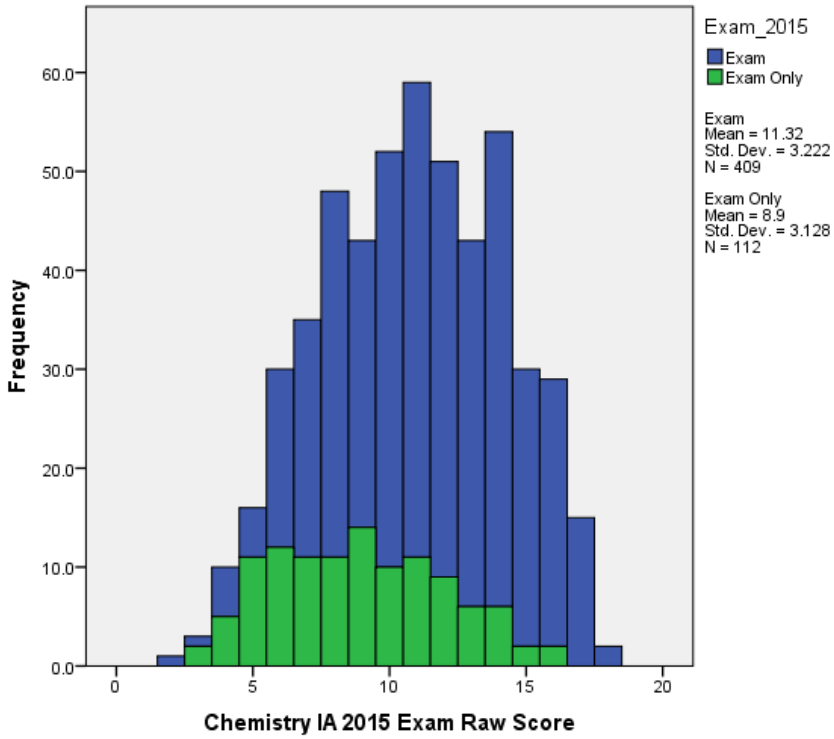

Figure 552: The Results of Students on Overlapping Items within Chemistry IA Redeemable Exam MCQ Assessment in 2015 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

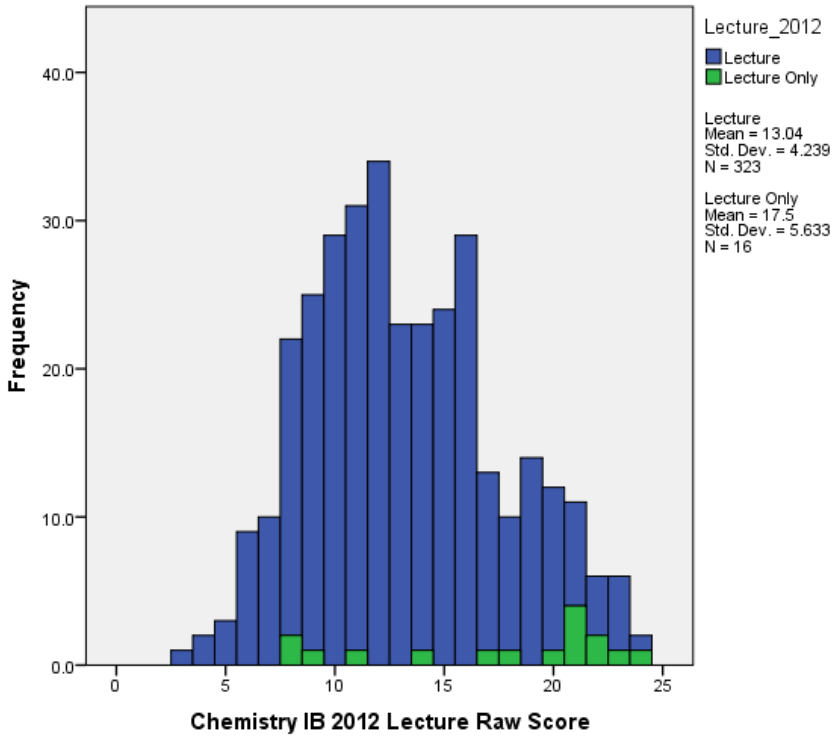

Figure 553: The Results of Students on Overlapping Items within Chemistry IB Lecture Test MCQ Assessments in 2012 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

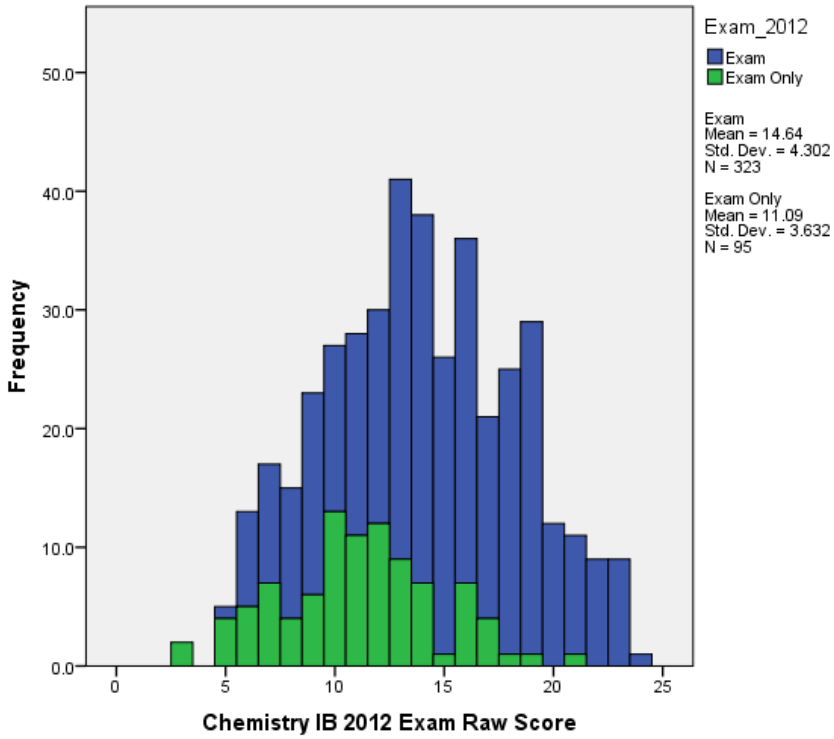

Figure 554: The Results of Students on Overlapping Items within Chemistry IB Redeemable Exam MCQ Assessment in 2012 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

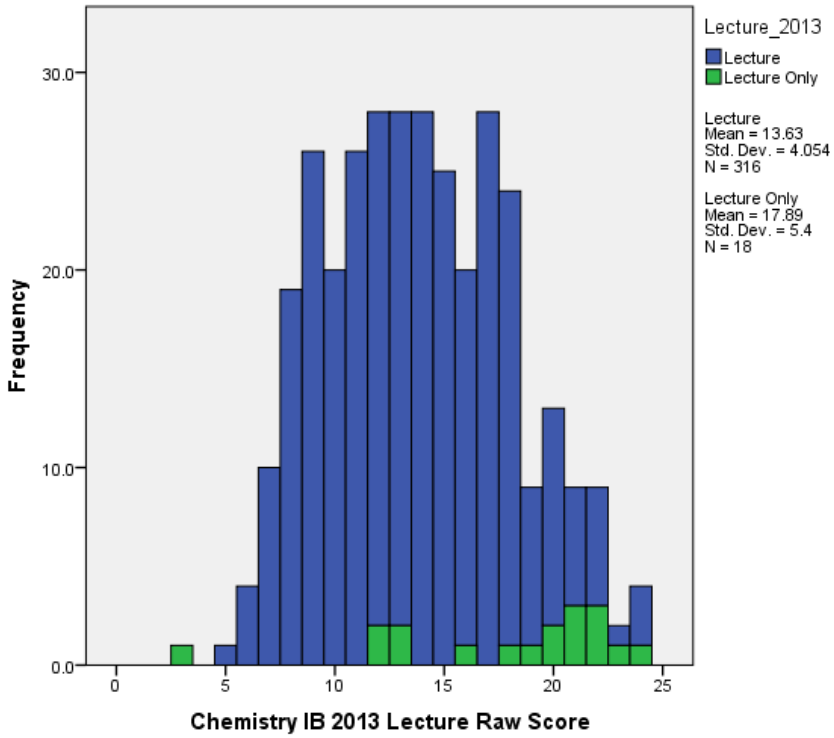

Figure 555: The Results of Students on Overlapping Items within Chemistry IB Lecture Test MCQ Assessments in 2013 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

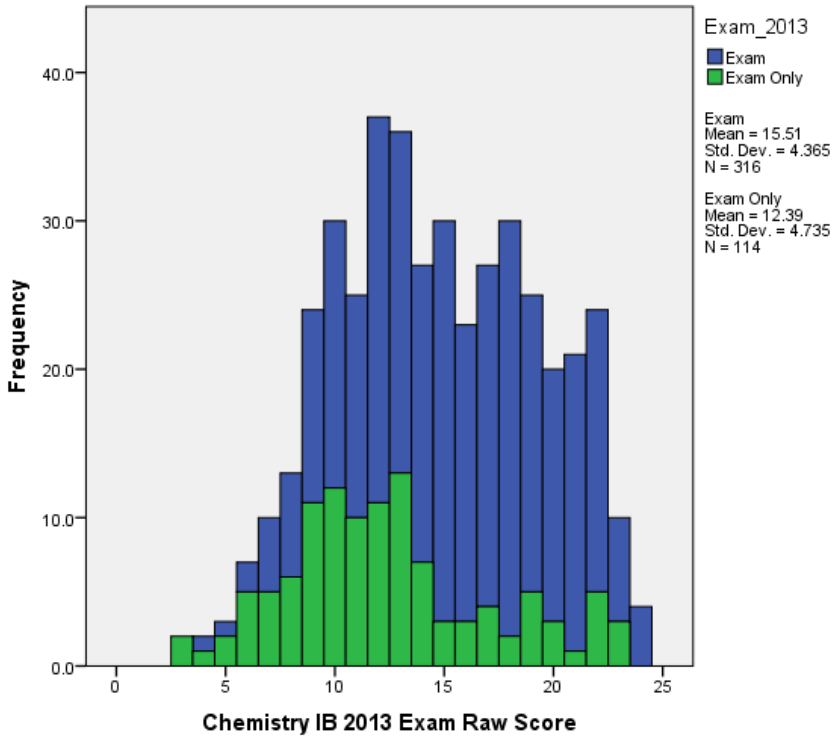

Figure 556: The Results of Students on Overlapping Items within Chemistry IB Redeemable Exam MCQ Assessment in 2013 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

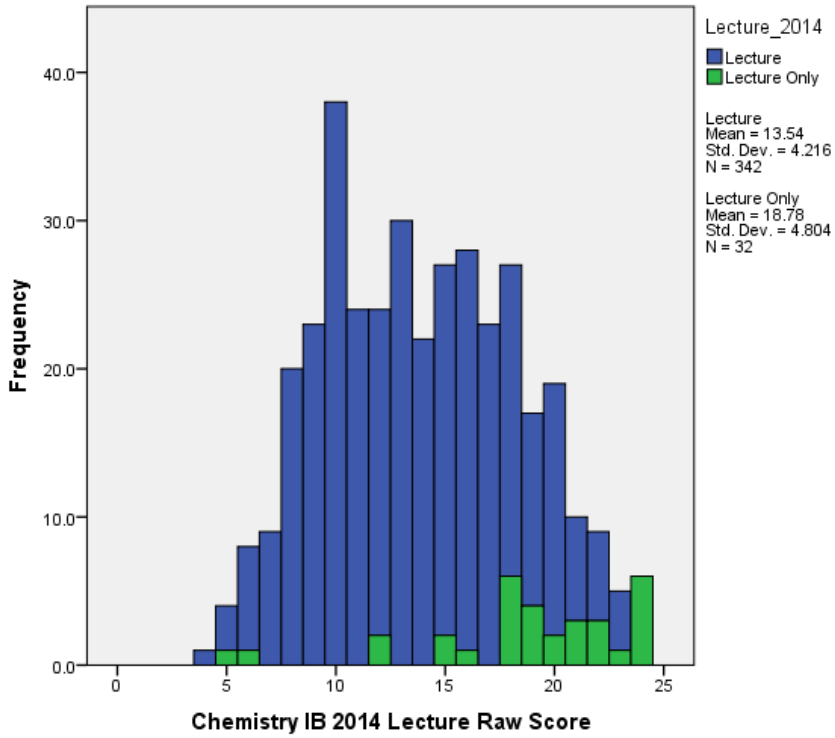

Figure 557: The Results of Students on Overlapping Items within Chemistry IB Lecture Test MCQ Assessments in 2014 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

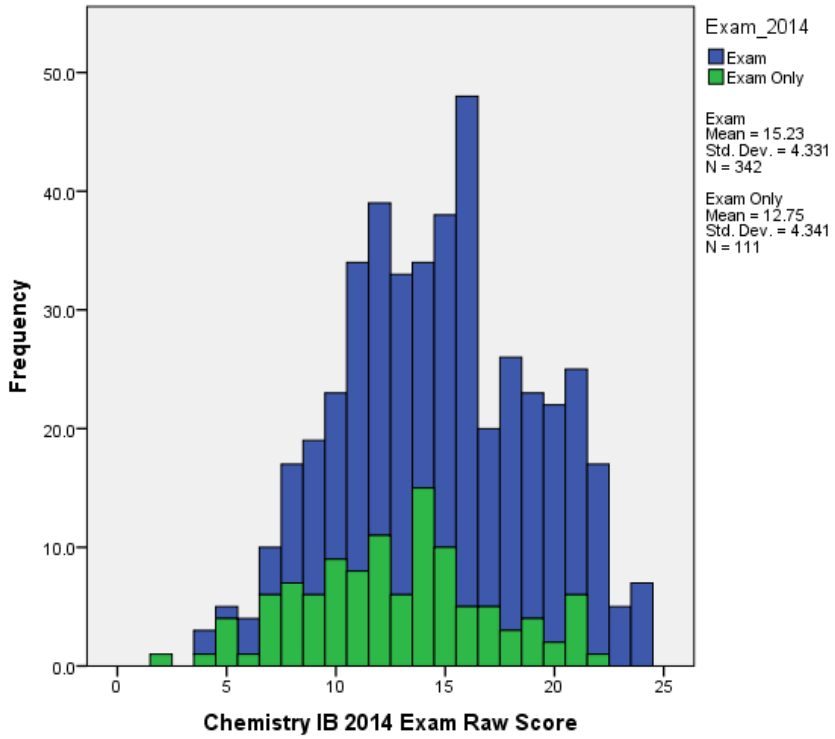

Figure 558: The Results of Students on Overlapping Items within Chemistry IB Redeemable Exam MCQ Assessment in 2014 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

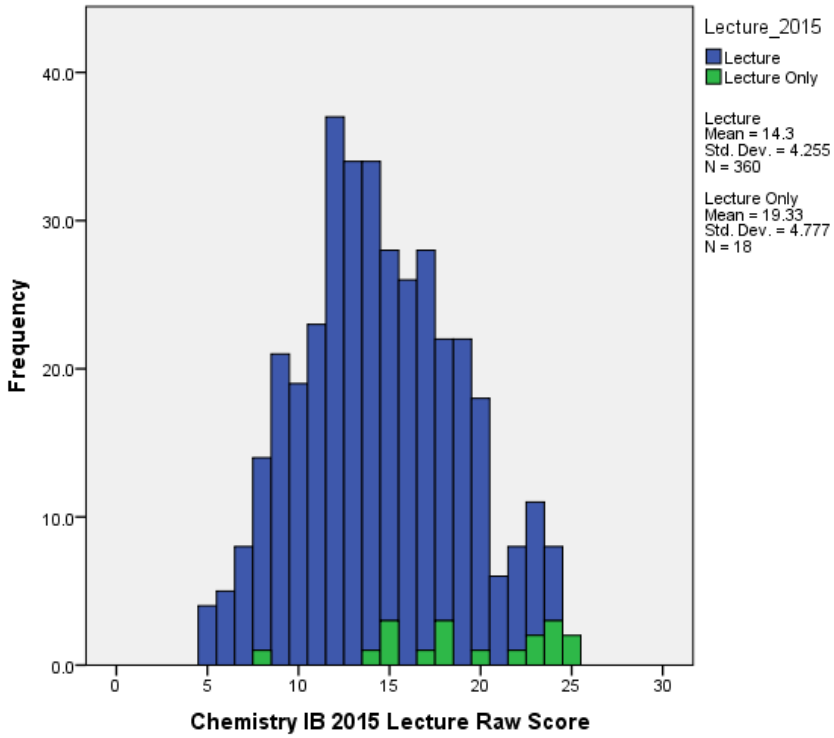

Figure 559: The Results of Students on Overlapping Items within Chemistry IB Lecture Test MCQ Assessments in 2015 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

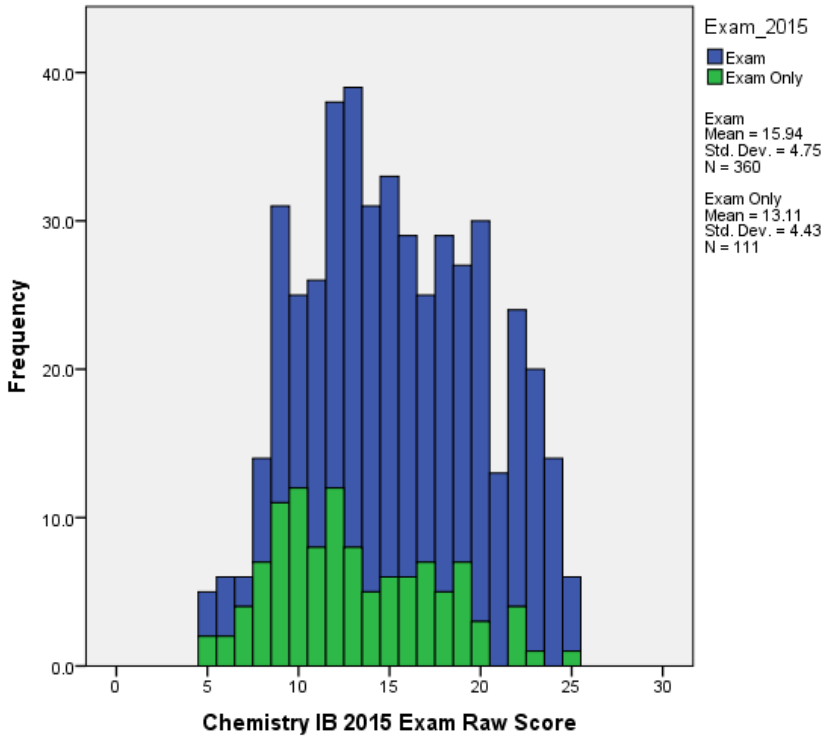

Figure 560: The Results of Students on Overlapping Items within Chemistry IB Redeemable Exam MCQ Assessment in 2015 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

7.21 Scatterplot Comparison of Student Ability Measures in Test-Retest MCQ Assessments

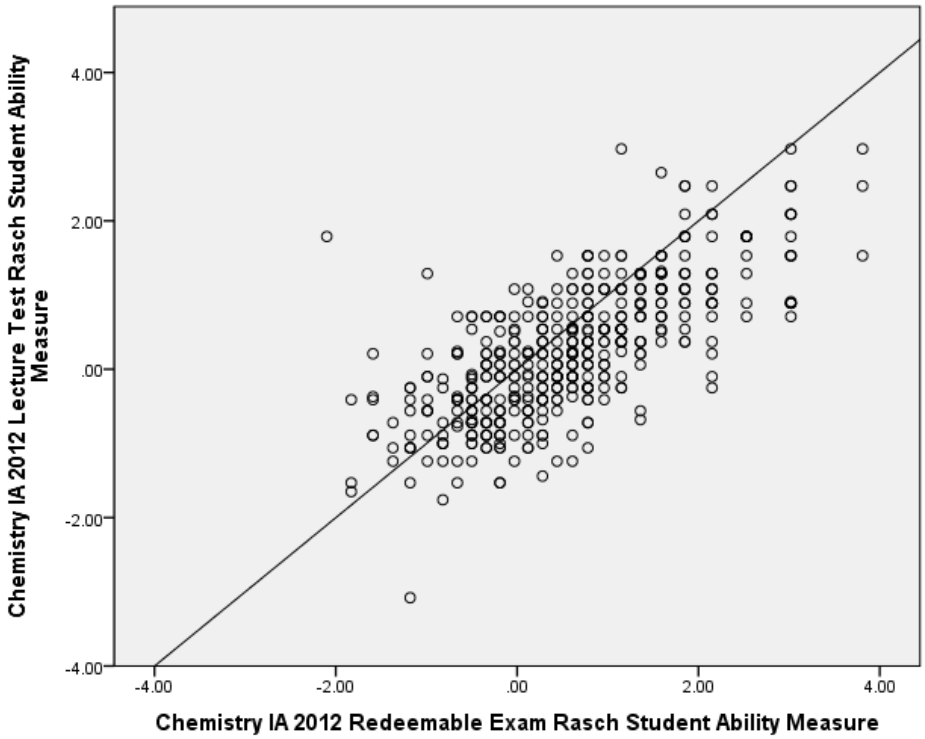

Figure 561: Scatterplot Comparison of Student Ability Measures Obtained Using the Same Items in Two Assessments within Chemistry IA during 2012 to View Changes in Student Performance

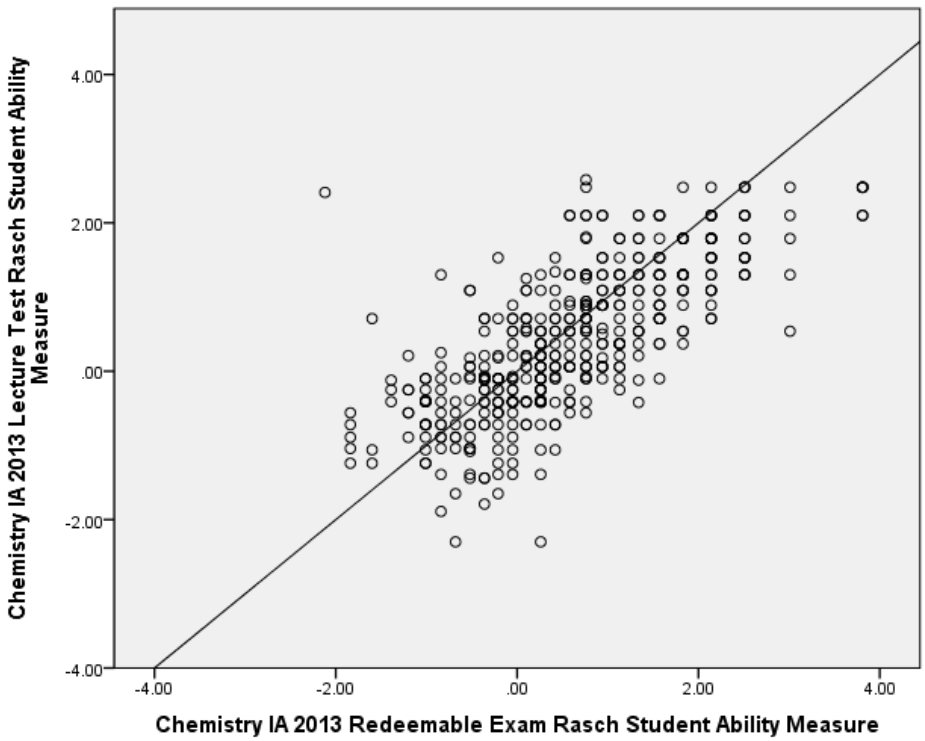

Figure 562: Scatterplot Comparison of Student Ability Measures Obtained Using the Same Items in Two Assessments within Chemistry IA during 2013 to View Changes in Student Performance

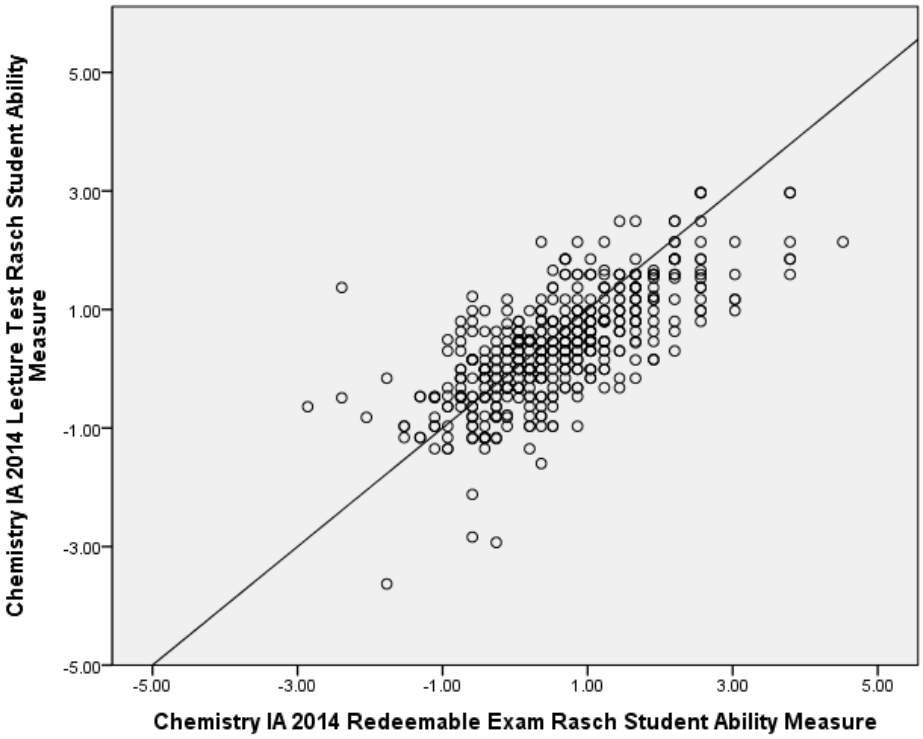

Figure 563: Scatterplot Comparison of Student Ability Measures Obtained Using the Same Items in Two Assessments within Chemistry IA during 2014 to View Changes in Student Performance

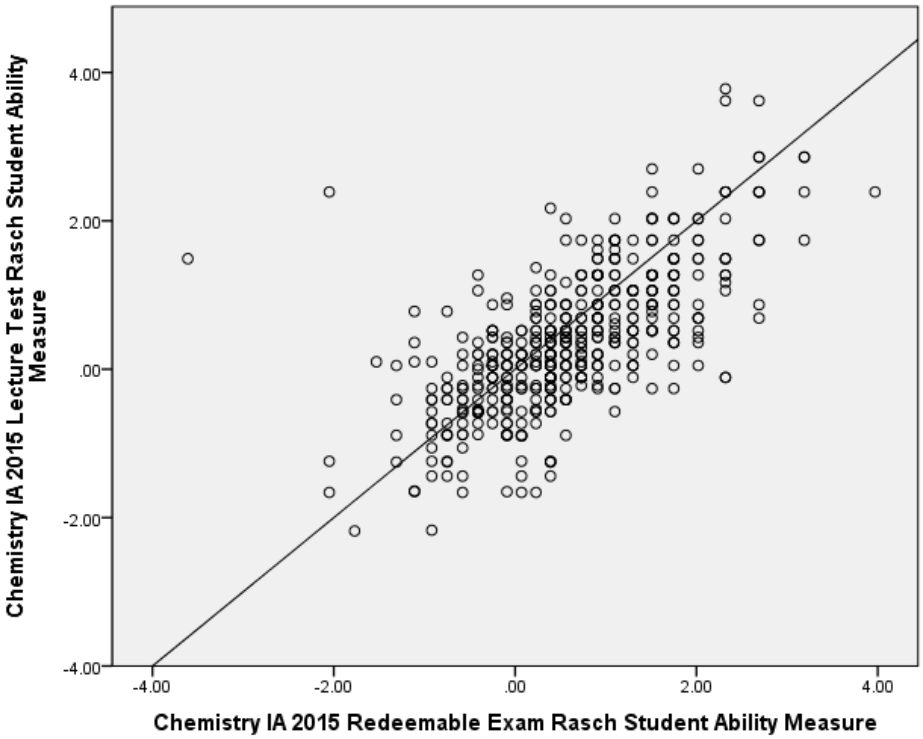

Figure 564: Scatterplot Comparison of Student Ability Measures Obtained Using the Same Items in Two Assessments within Chemistry IA during 2015 to View Changes in Student Performance

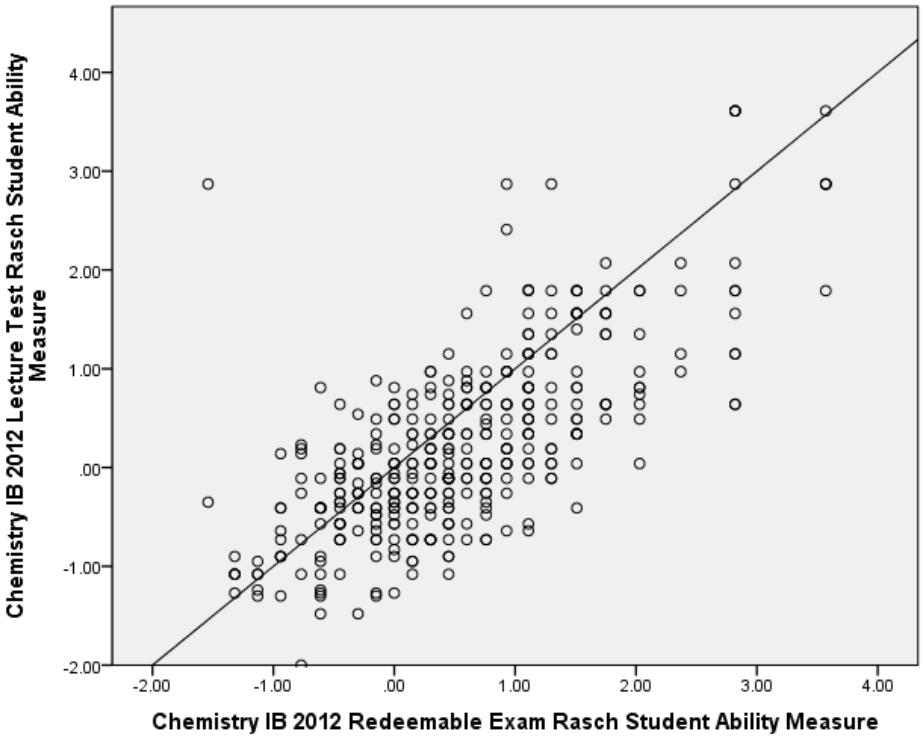

Figure 565: Scatterplot Comparison of Student Ability Measures Obtained Using the Same Items in Two Assessments within Chemistry IB during 2012 to View Changes in Student Performance

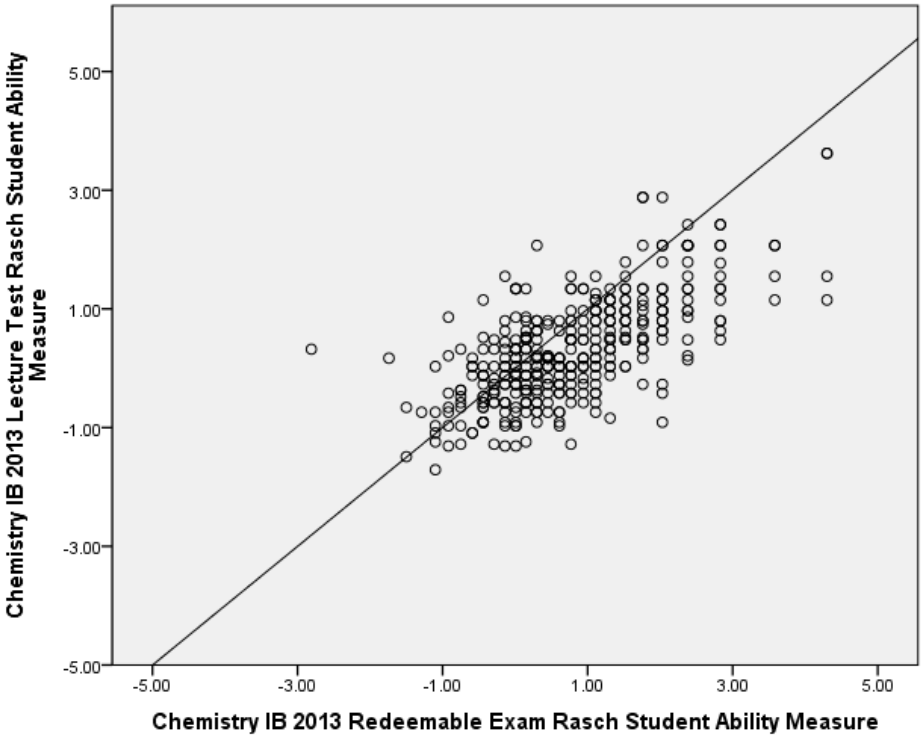

Figure 566: Scatterplot Comparison of Student Ability Measures Obtained Using the Same Items in Two Assessments within Chemistry IB during 2013 to View Changes in Student Performance

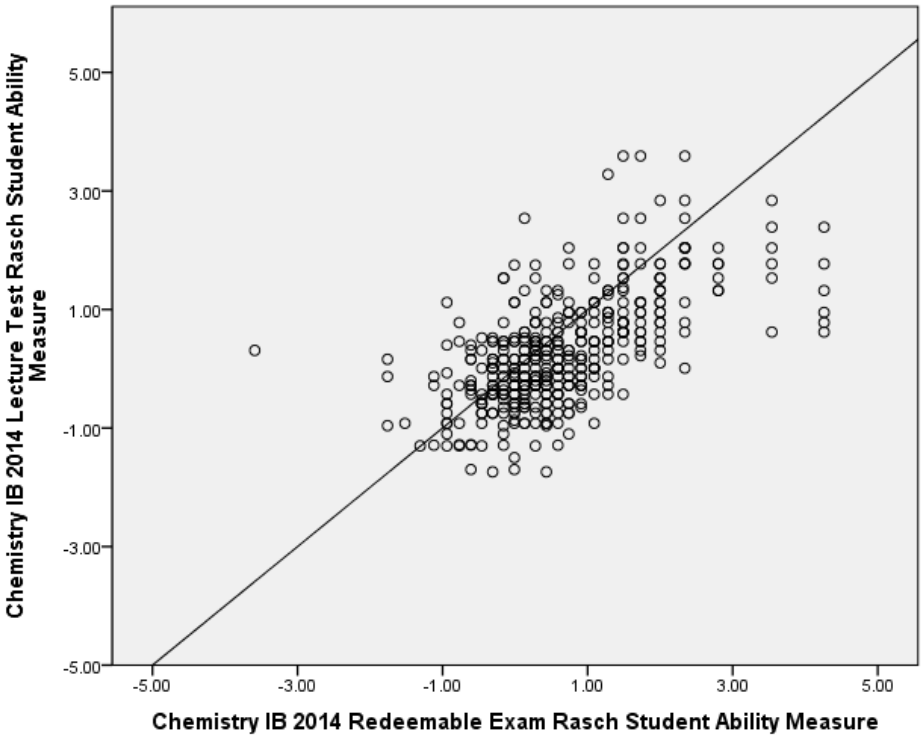

Figure 567: Scatterplot Comparison of Student Ability Measures Obtained Using the Same Items in Two Assessments within Chemistry IB during 2014 to View Changes in Student Performance

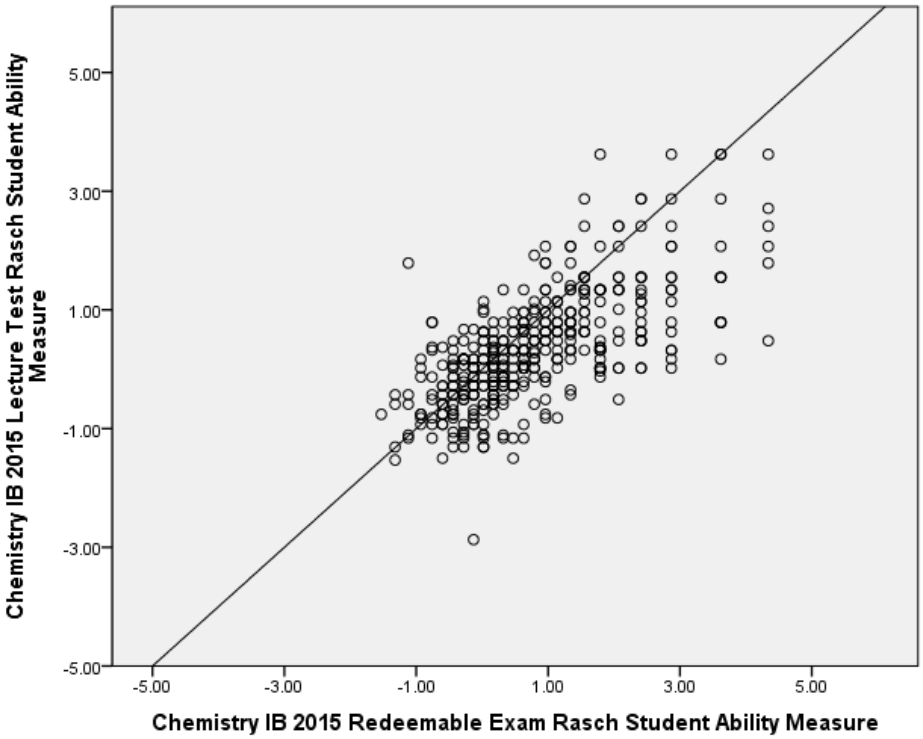

Figure 568: Scatterplot Comparison of Student Ability Measures Obtained Using the Same Items in Two Assessments within Chemistry IB during 2015 to View Changes in Student Performance

7.22 Distribution Comparison of Student Ability Measures in Test-Retest MCQ Assessments

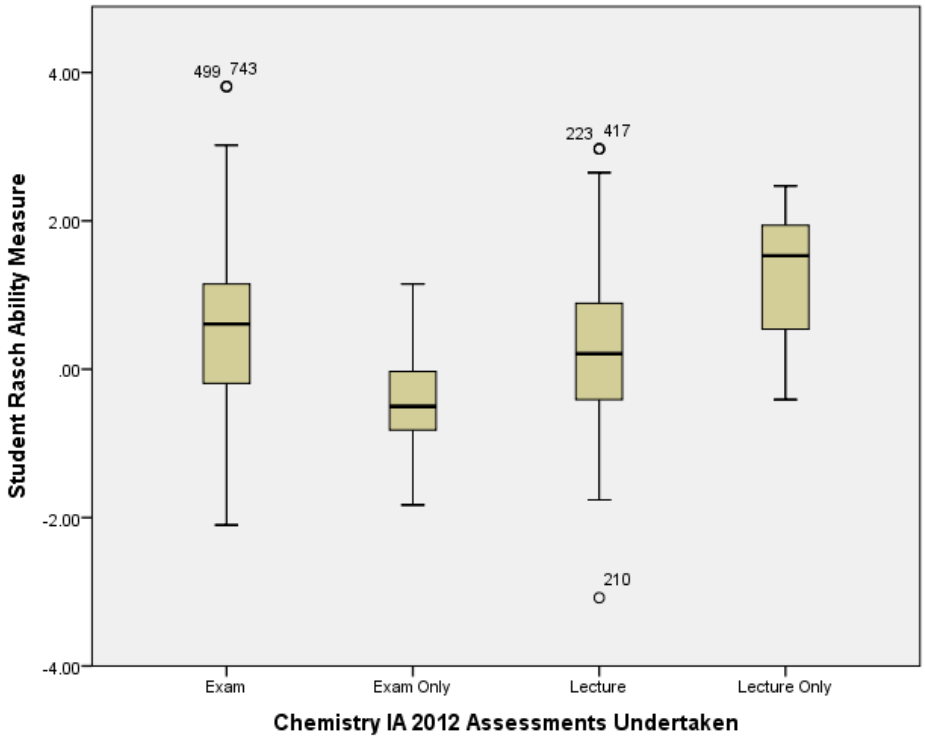

Figure 569: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IA During 2012, Separating Students Who Only Undertook One Assessment

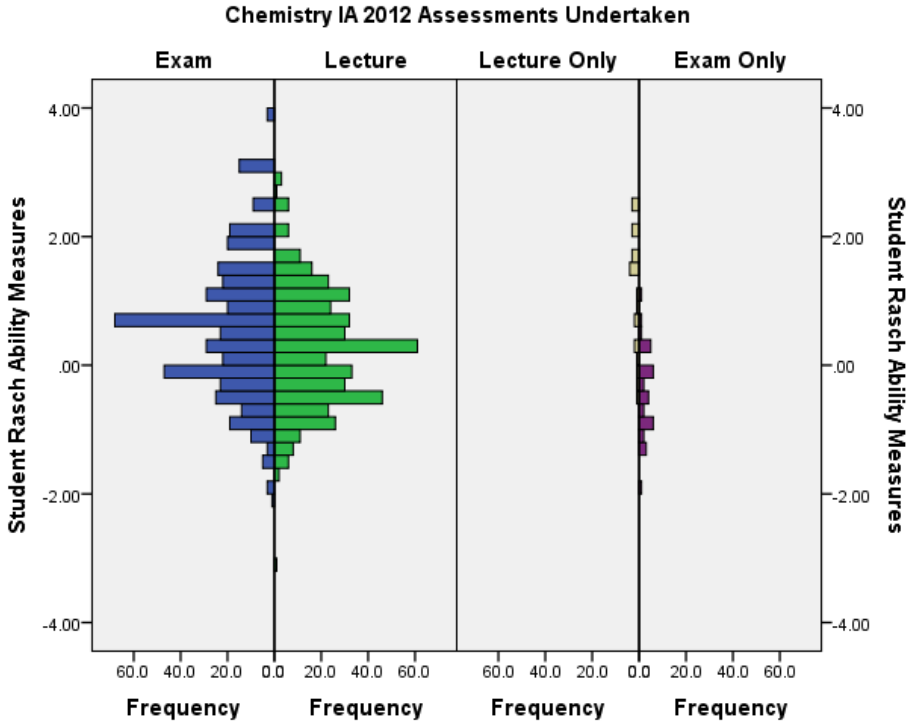

Figure 570: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA During 2012, Separating Student Who Only Undertook One Assessment

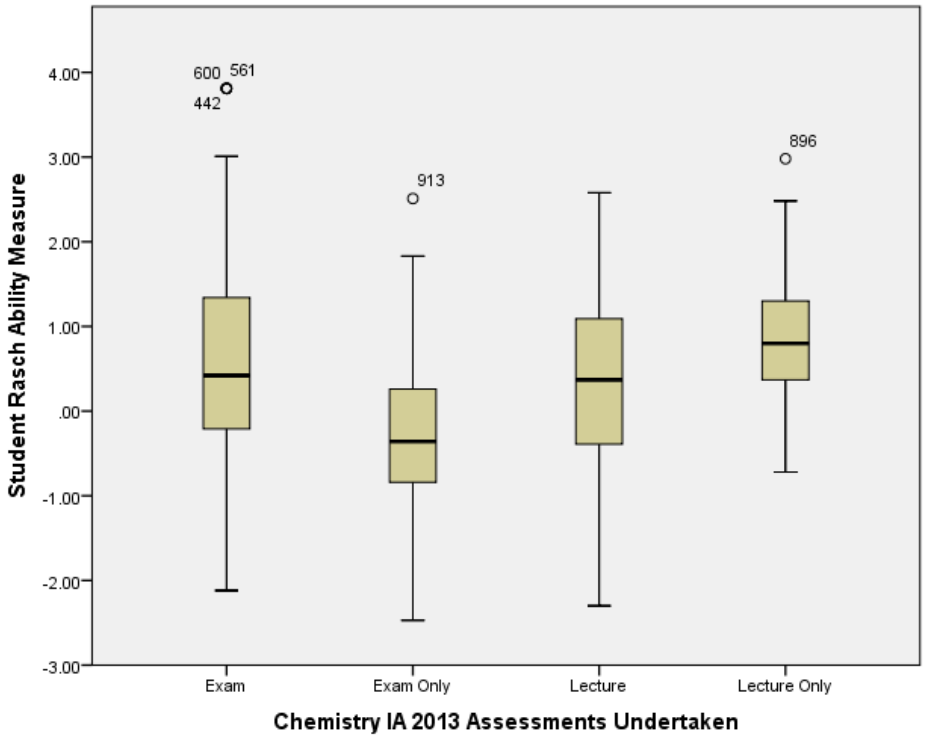

Figure 571: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IA During 2013, Separating Students Who Only Undertook One Assessment

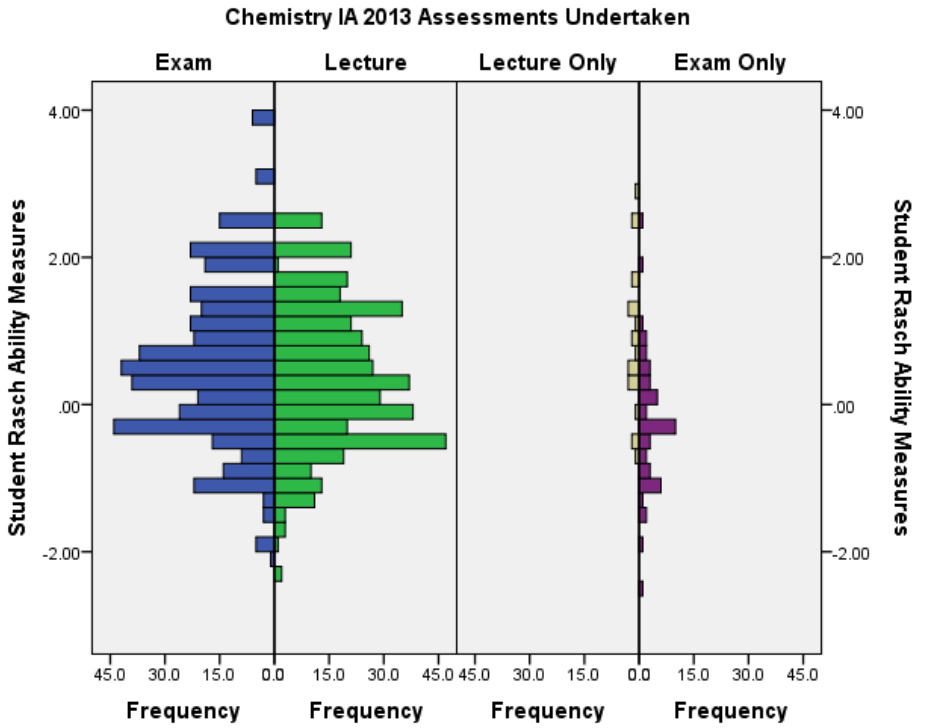

Figure 572: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA During 2013, Separating Student Who Only Undertook One Assessment

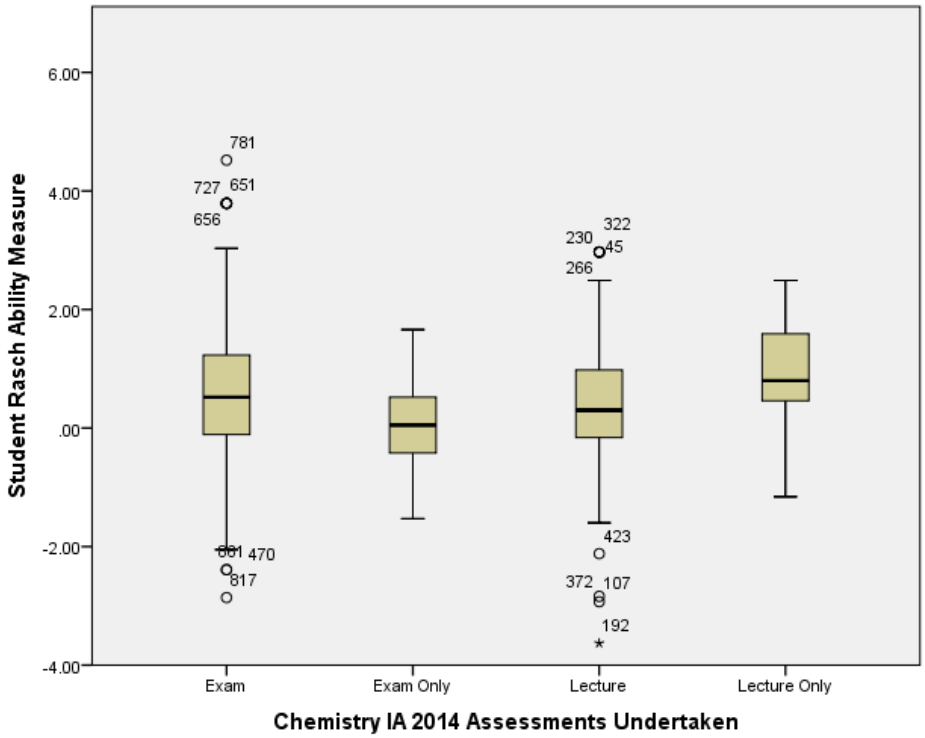

Figure 573: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IA During 2014, Separating Students Who Only Undertook One Assessment

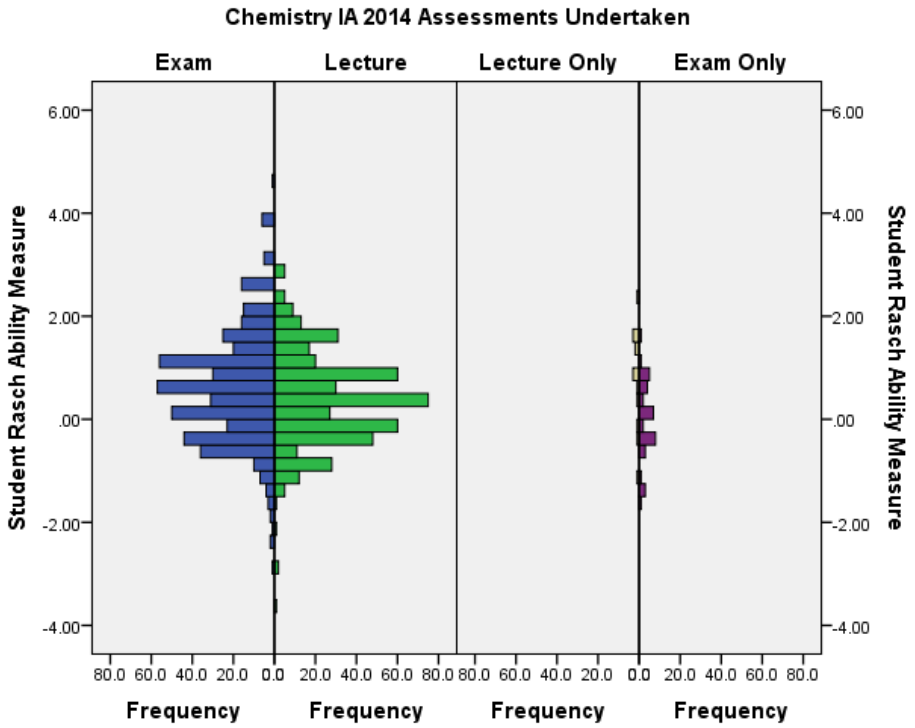

Figure 574: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA During 2014, Separating Student Who Only Undertook One Assessment

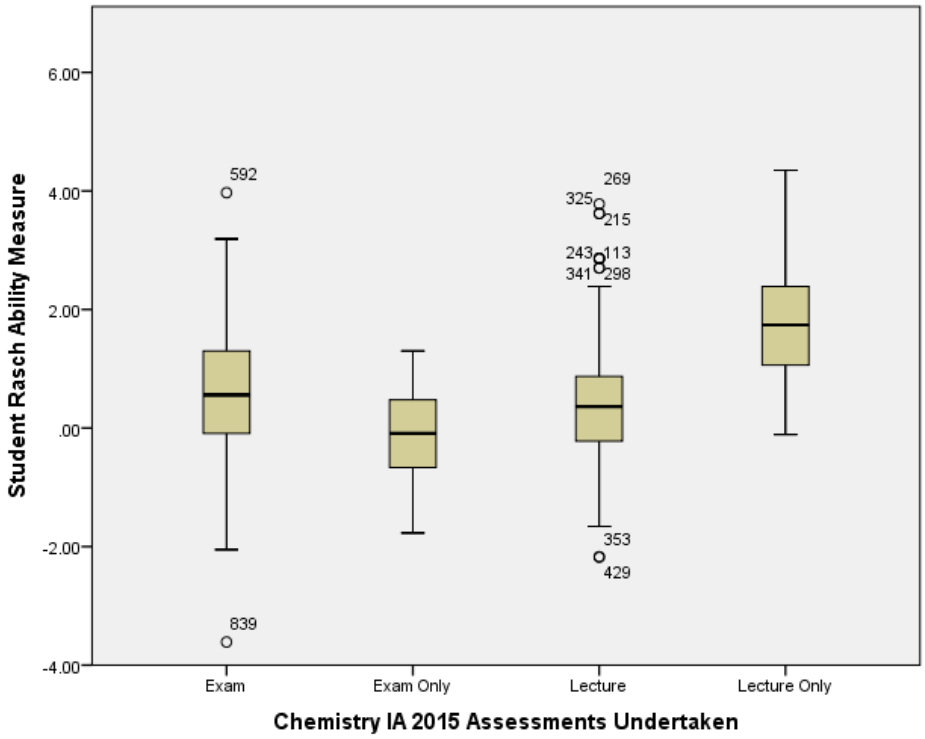

Figure 575: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IA During 2015, Separating Students Who Only Undertook One Assessment

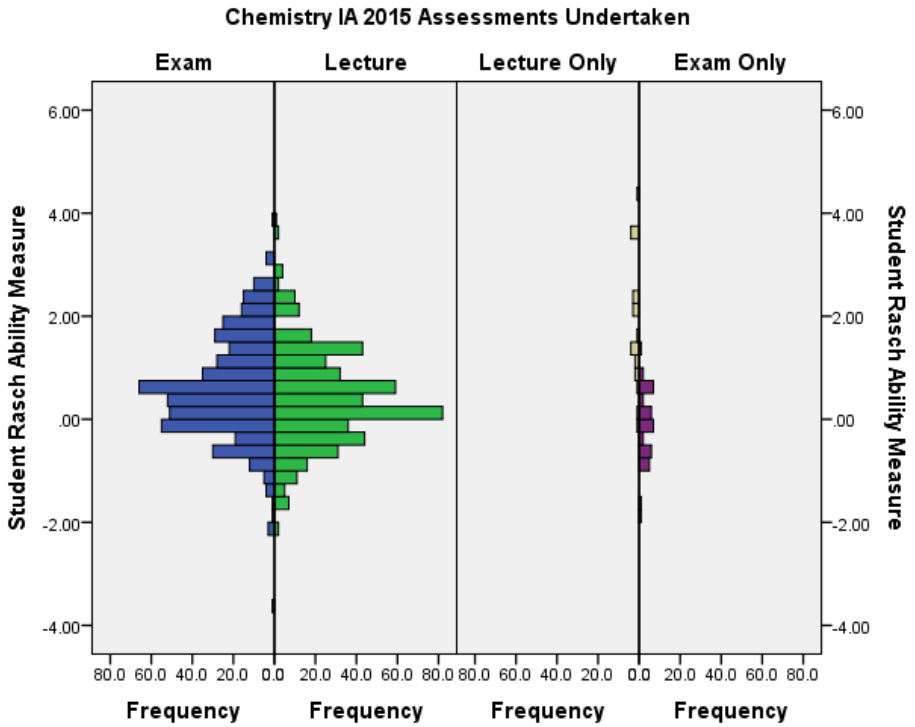

Figure 576: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IA During 2015, Separating Student Who Only Undertook One Assessment

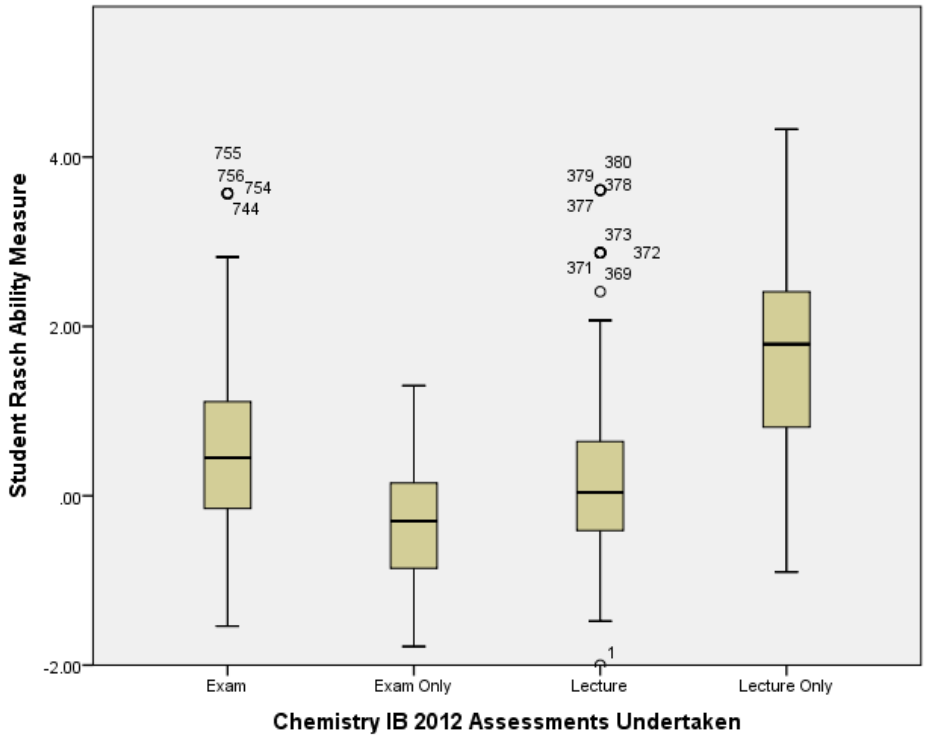

Figure 577: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IB During 2012, Separating Students Who Only Undertook One Assessment

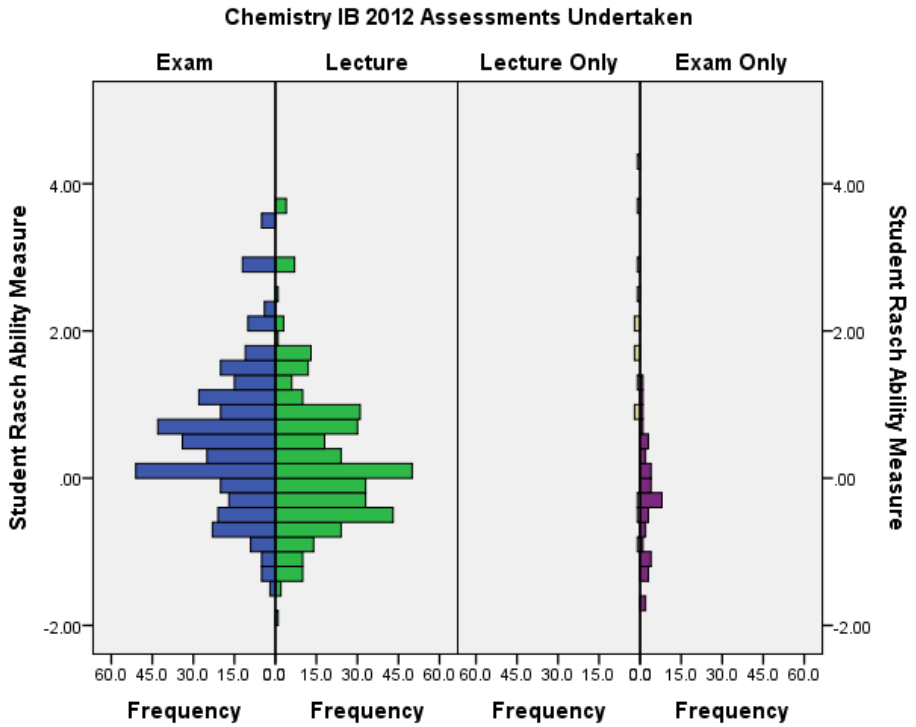

Figure 578: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IB During 2012, Separating Student Who Only Undertook One Assessment

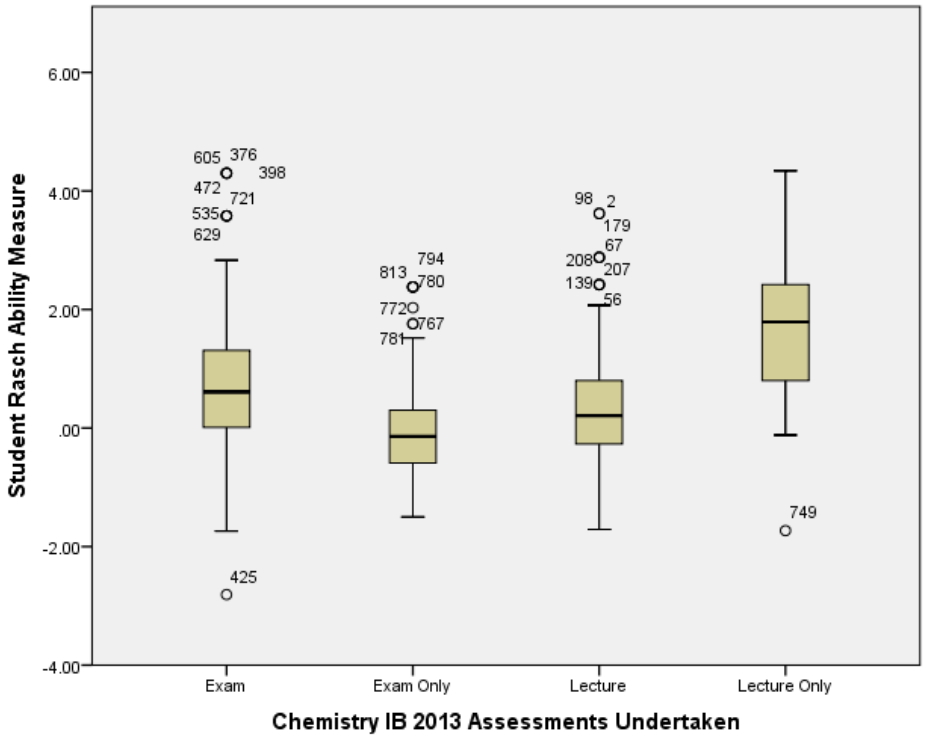

Figure 579: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IB During 2013, Separating Students Who Only Undertook One Assessment

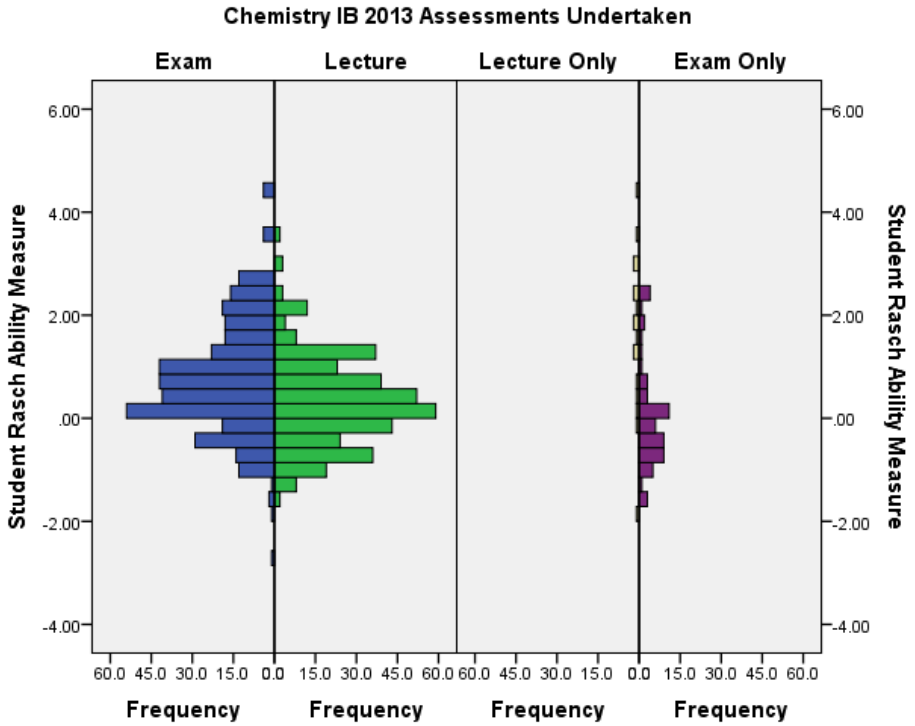

Figure 580: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IB During 2013, Separating Student Who Only Undertook One Assessment

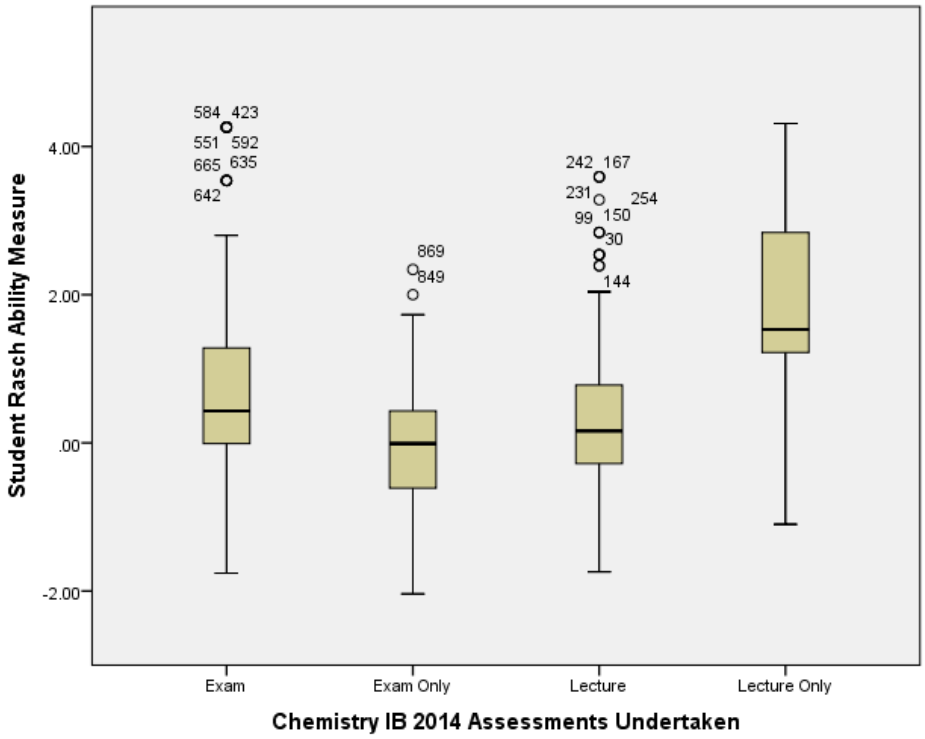

Figure 581: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IB During 2014, Separating Students Who Only Undertook One Assessment

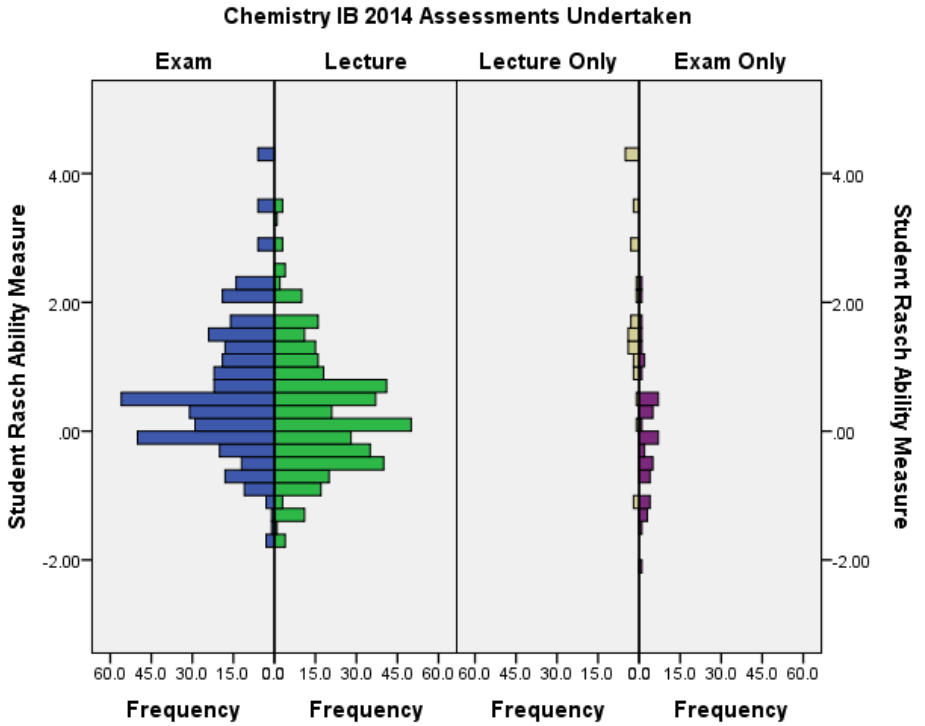

Figure 582: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IB During 2014, Separating Student Who Only Undertook One Assessment

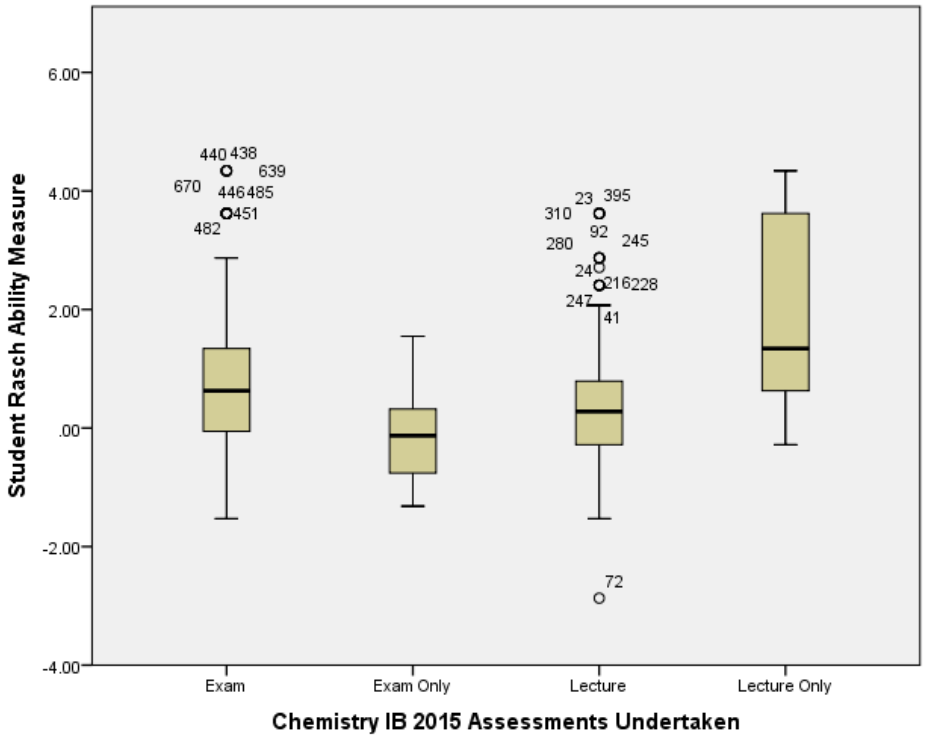

Figure 583: Boxplot Distribution of Student Raw Scores in Items Shared Across Assessments in Chemistry IB During 2015, Separating Students Who Only Undertook One Assessment

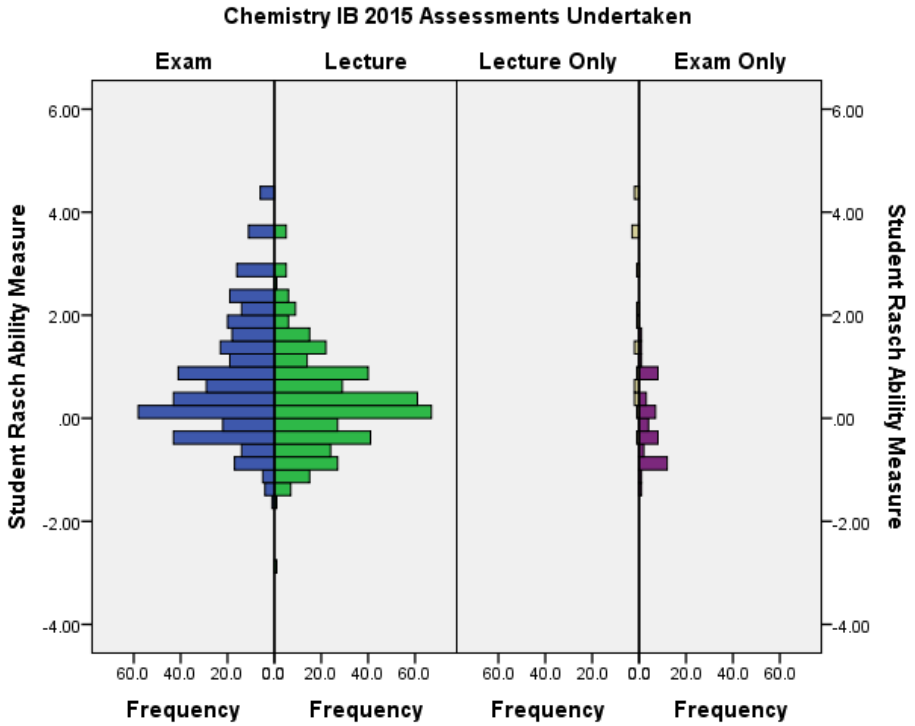

Figure 584: Histogram Distribution of Student Raw Scores in Items Shared Across Both Assessments in Chemistry IB During 2015, Separating Student Who Only Undertook One Assessment

7.23 Comparison of Changes in Student Ability Measures in Shared Items

Table 93: Ability Measure Average Result from Overlapping Items within Chemistry IA MCQ Assessments from 2012 Separating Students Based on Their Shift in Performance

Chemistry IA 2012							
		Count	Lecture Test Average Ability	Lecture Test S.D.	Exam Average Ability	Exam S.D.	Average Ability Change
Student Cohorts	Cohort	511	0.26	0.90	0.58	1.07	
	Lecture Only	24	1.16	0.96			
	Exam Only	34			-0.43	0.66	
Test-Retest	Increase	318	0.21	0.91	0.90	1.00	0.69
	Decrease	135	0.37	0.85	-0.17	0.85	-0.54

Table 94: Ability Measure Average Result from Overlapping Items within Chemistry IA MCQ Assessments from 2013 Separating Students Based on Their Shift in Performance

Chemistry IA 2013							
		Count	Lecture Test Average Ability	Lecture Test S.D.	Exam Average Ability	Exam S.D.	Average Ability Change
Student Cohorts	Cohort	510	0.41	0.99	0.56	1.08	
	Lecture Only	22	0.89	0.97			
	Exam Only	49			-0.26	0.91	
Test-Retest	Increase	277	0.31	1.02	0.90	1.04	0.60
	Decrease	162	0.59	0.91	-0.02	0.89	-0.61

Table 95: Ability Measure Average Result from Overlapping Items within Chemistry IA MCQ Assessments from 2014 Separating Students Based on Their Shift in Performance

Chemistry IA 2014							
		Count	Lecture Test Average Ability	Lecture Test S.D.	Exam Average Ability	Exam S.D.	Average Ability Change
Student Cohorts	Cohort	513	0.39	0.93	0.60	1.08	
	Lecture Only	14	0.86	0.90			
	Exam Only	38			-0.01	0.74	
Test-Retest	Increase	294	0.30	0.94	0.95	1.00	0.65
	Decrease	167	0.56	0.88	-0.02	0.92	-0.57

Table 96: Ability Measure Average Result from Overlapping Items within Chemistry IA MCQ Assessments from 2015 Separating Students Based on Their Shift in Performance

Chemistry IA 2015							
		Count	Lecture Test Average Ability	Lecture Test S.D.	Exam Average Ability	Exam S.D.	Average Ability Change
Student Cohorts	Cohort	548	0.39	0.93	0.60	0.97	
	Lecture Only	23	1.91	1.19			
	Exam Only	40			-0.11	0.69	
Test-Retest	Increase	311	0.19	0.84	0.81	0.89	0.62
	Decrease	173	0.76	0.97	0.21	0.99	-0.54
	No Change	1	-0.41	0.00	-0.41	0.00	

Table 97: Ability Measure Average Result from Overlapping Items within Chemistry IB MCQ Assessments from 2012 Separating Students Based on Their Shift in Performance

Chemistry IB 2012							
		Count	Lecture Test Average Ability	Lecture Test S.D.	Exam Average Ability	Exam S.D.	Average Ability Change
Student Cohorts	Cohort	434	0.22	0.92	0.52	0.96	
	Lecture Only	14	1.60	1.46			
	Exam Only	40			-0.31	0.73	
Test-Retest	Increase	242	0.03	0.80	0.73	0.89	0.70
	Decrease	138	0.55	1.03	0.17	0.96	-0.39

Table 98: Ability Measure Average Result from Overlapping Items within Chemistry IB MCQ Assessments from 2013 Separating Students Based on Their Shift in Performance

Chemistry IB 2013							
		Count	Lecture Test Average Ability	Lecture Test S.D.	Exam Average Ability	Exam S.D.	Average Ability Change
Student Cohorts	Cohort	451	0.32	0.87	0.75	1.08	
	Lecture Only	17	1.62	1.45			
	Exam Only	60			0.04	0.99	
Test-Retest	Increase	264	0.22	0.87	1.04	1.04	0.82
	Decrease	110	0.55	0.83	0.06	0.86	-0.49

Table 99: Ability Measure Average Result from Overlapping Items within Chemistry IB MCQ Assessments from 2014 Separating Students Based on Their Shift in Performance

Chemistry IB 2014							
		Count	Lecture Test Average Ability	Lecture Test S.D.	Exam Average Ability	Exam S.D.	Average Ability Change
Student Cohorts	Cohort	486	0.30	0.94	0.66	1.05	
	Lecture Only	31	1.96	1.46			
	Exam Only	48			-0.02	0.93	
Test-Retest	Increase	271	0.10	0.86	0.91	1.05	0.81
	Decrease	136	0.71	0.95	0.16	0.85	-0.55

Table 100: Ability Measure Average Result from Overlapping Items within Chemistry IB MCQ Assessments from 2015 Separating Students Based on Their Shift in Performance

Chemistry IB 2015							
		Count	Lecture Test Average Ability	Lecture Test S.D.	Exam Average Ability	Exam S.D.	Average Ability Change
Student Cohorts	Cohort	489	0.34	0.94	0.77	1.16	
	Lecture Only	17	1.86	1.52			
	Exam Only	49			-0.08	0.72	
Test-Retest	Increase	273	0.16	0.87	1.05	1.16	0.89
	Decrease	122	0.56	0.94	0.07	0.85	-0.49
	No Change	28	1.09	0.01	1.09	0.01	

7.24 Histogram Comparison of Student Ability in Test-Retest Assessments

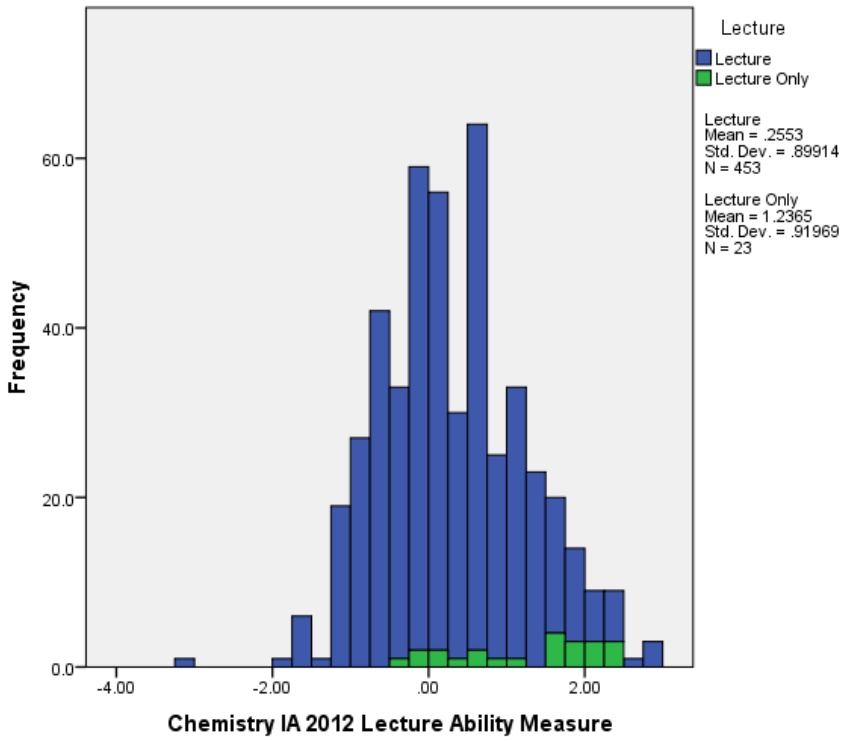

Figure 585: The Student Ability of Students on Overlapping Items within Chemistry IA Lecture Test MCQ Assessments in 2012 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

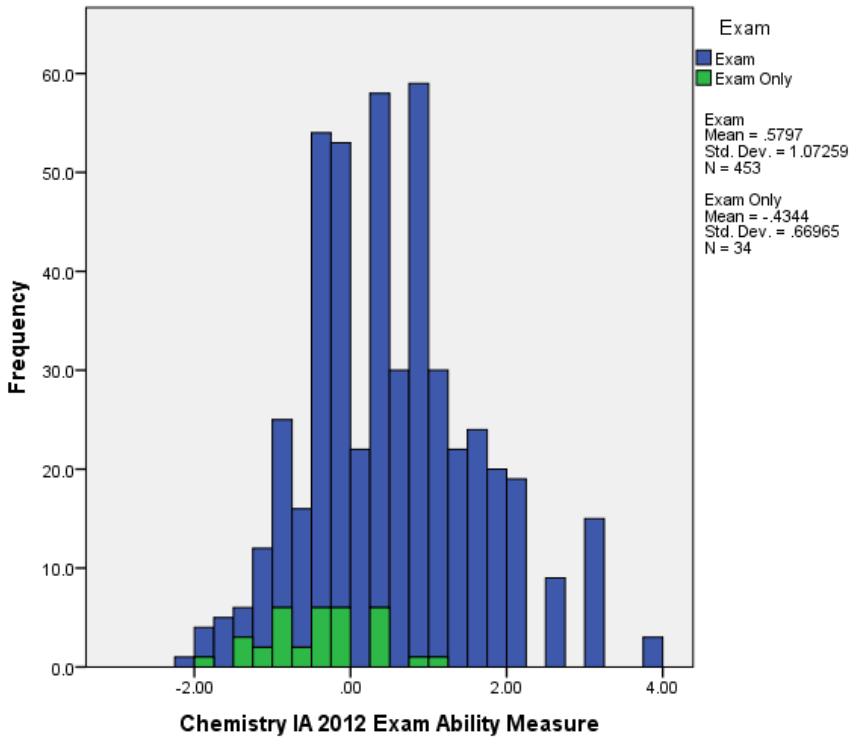

Figure 586: The Student Ability of Students on Overlapping Items within Chemistry IA Redeemable Exam MCQ Assessments in 2012 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

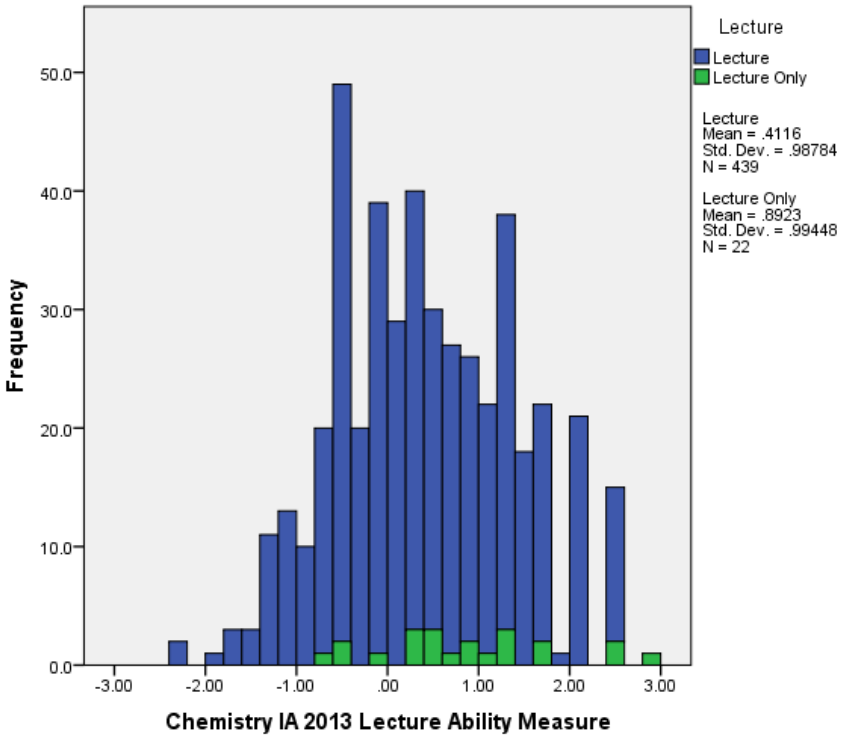

Figure 587: The Student Ability of Students on Overlapping Items within Chemistry IA Lecture Test MCQ Assessments in 2013 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

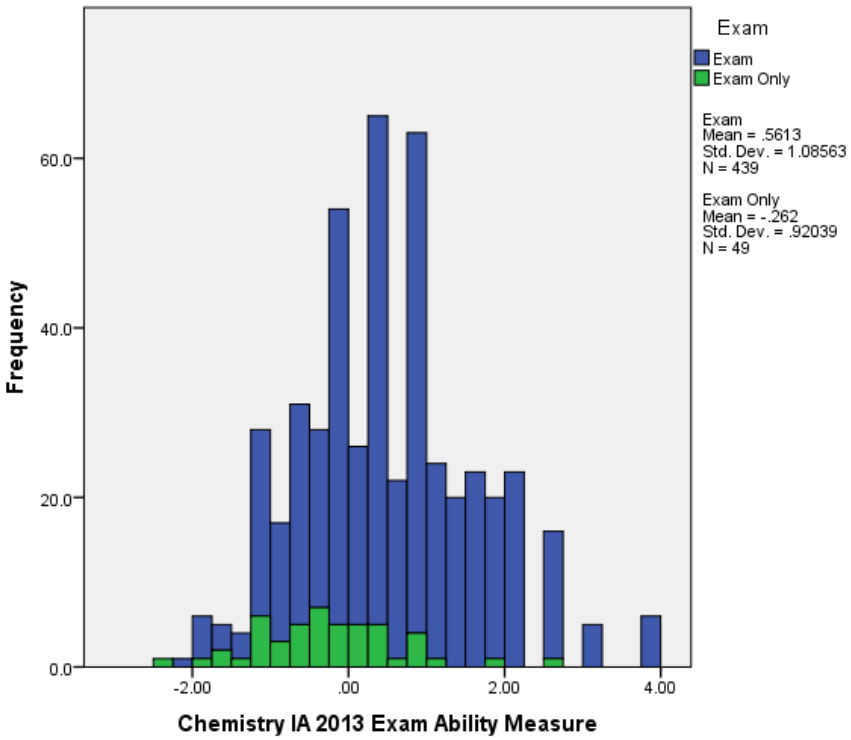

Figure 588: The Student Ability of Students on Overlapping Items within Chemistry IA Redeemable Exam MCQ Assessments in 2013 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

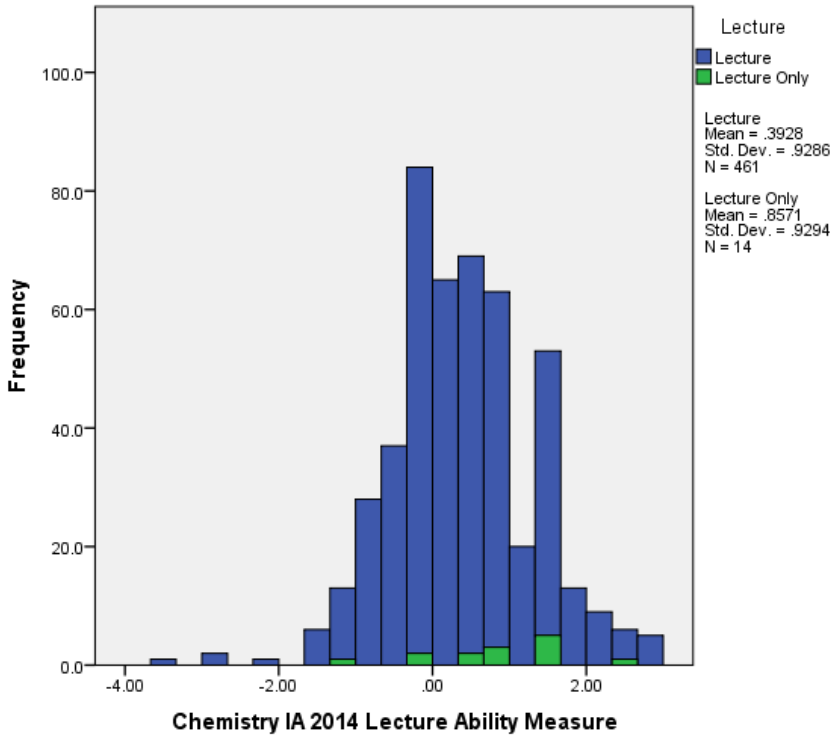

Figure 589: The Student Ability of Students on Overlapping Items within Chemistry IA Lecture Test MCQ Assessments in 2014 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

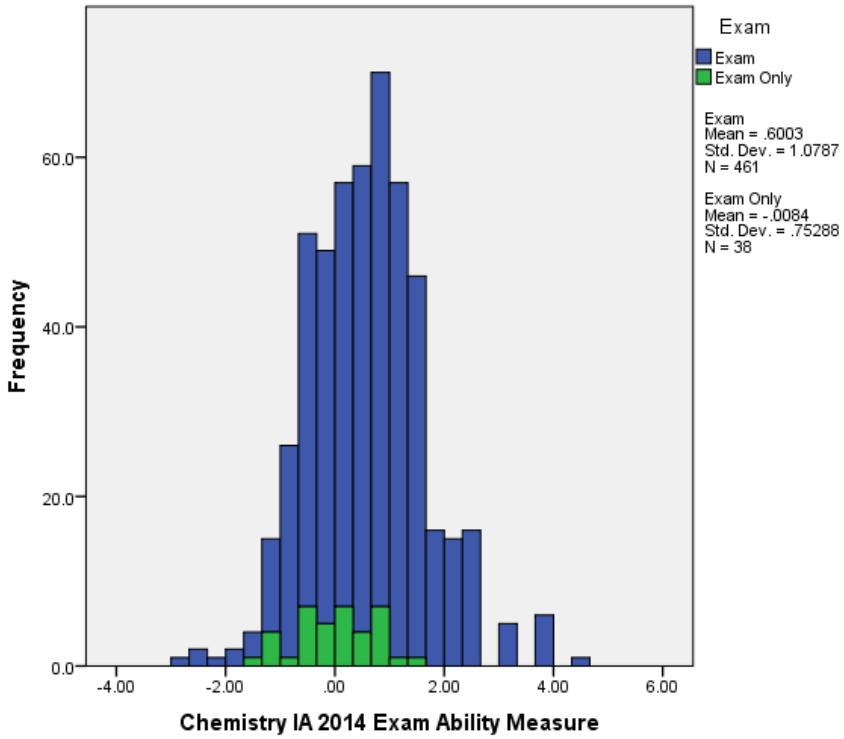

Figure 590: The Student Ability of Students on Overlapping Items within Chemistry IA Redeemable Exam MCQ Assessments in 2014 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

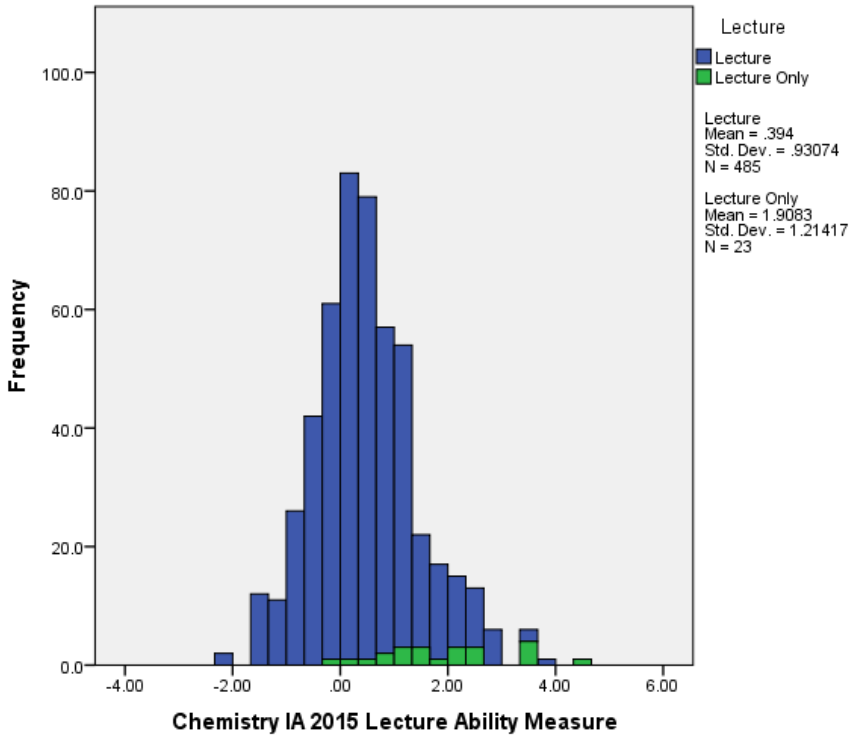

Figure 591: The Student Ability of Students on Overlapping Items within Chemistry IA Lecture Test MCQ Assessments in 2015 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

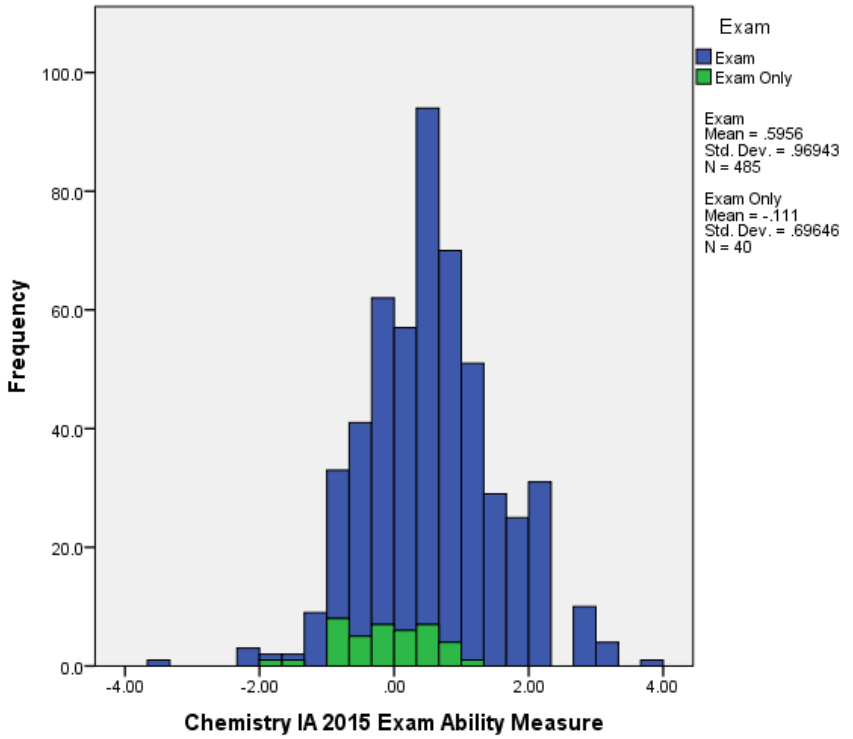

Figure 592: The Student Ability of Students on Overlapping Items within Chemistry IA Redeemable Exam MCQ Assessments in 2015 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

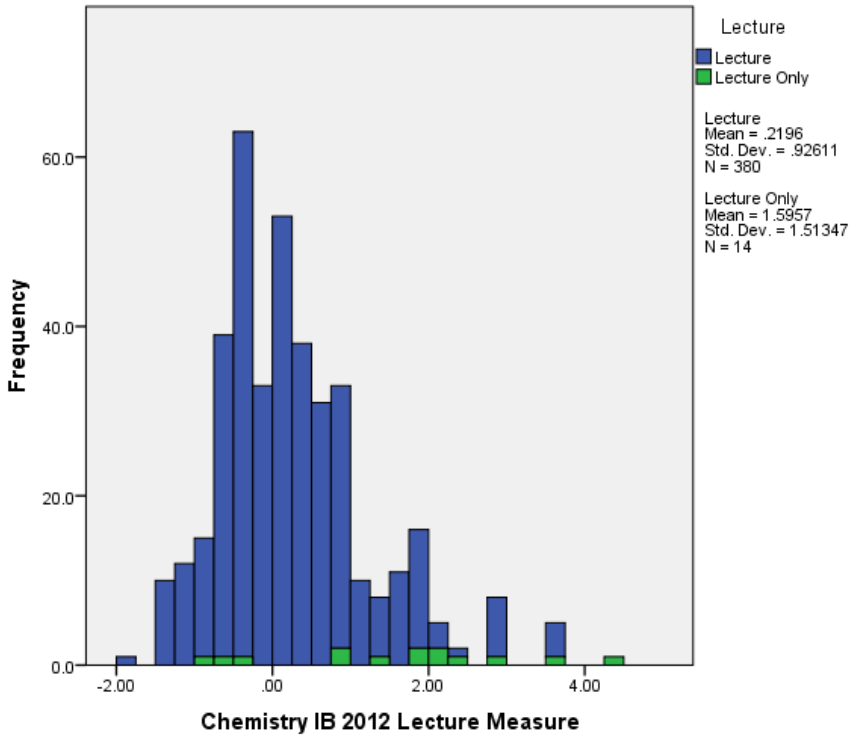

Figure 593: The Student Ability of Students on Overlapping Items within Chemistry IB Lecture Test MCQ Assessments in 2012 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

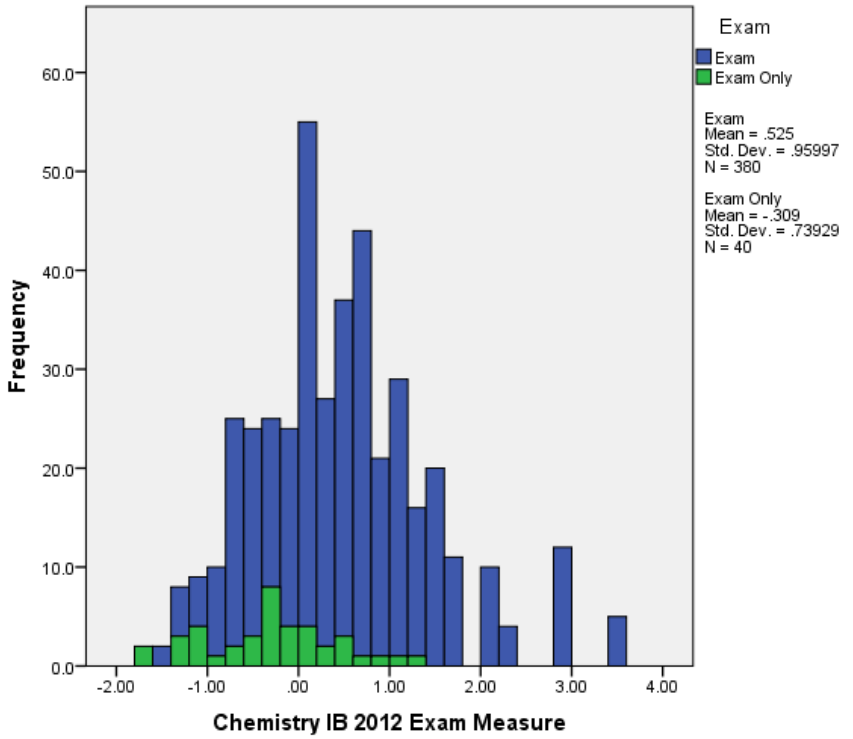

Figure 594: The Student Ability of Students on Overlapping Items within Chemistry IB Redeemable Exam MCQ Assessments in 2012 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

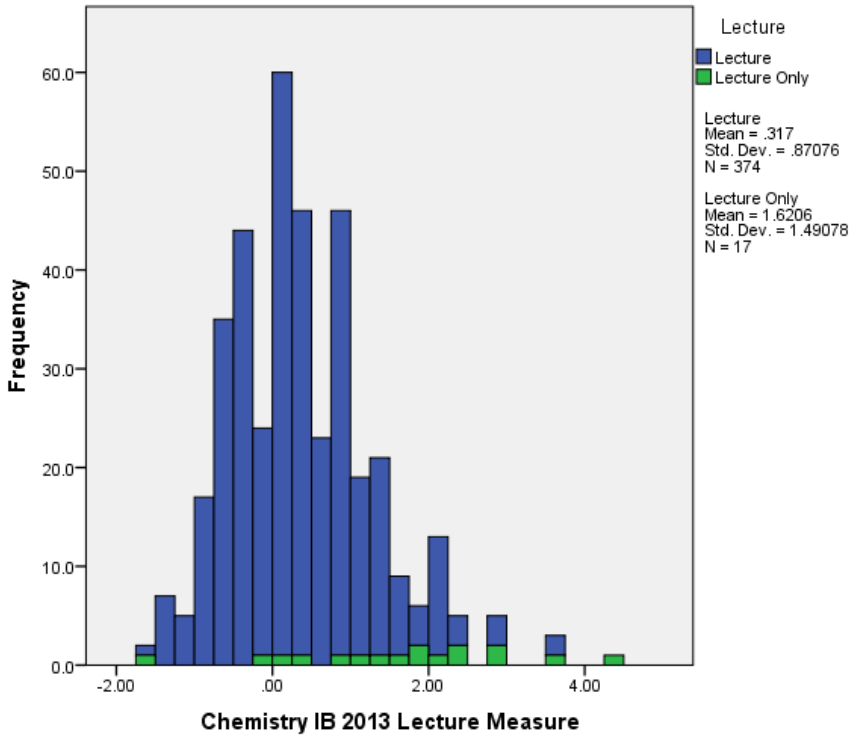

Figure 595: The Student Ability of Students on Overlapping Items within Chemistry IB Lecture Test MCQ Assessments in 2013 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

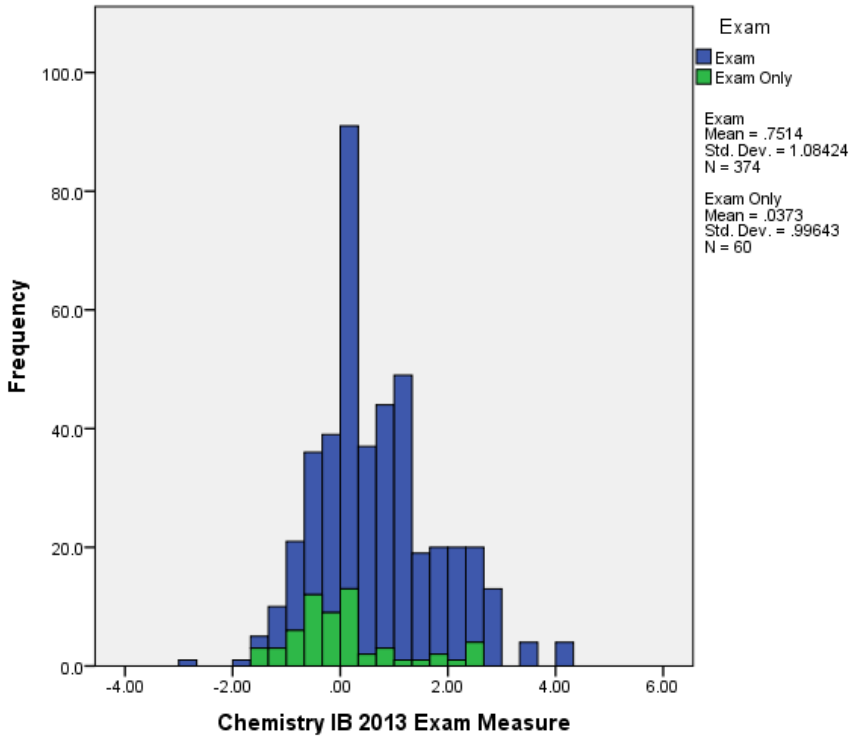

Figure 596: The Student Ability of Students on Overlapping Items within Chemistry IB Redeemable Exam MCQ Assessments in 2013 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

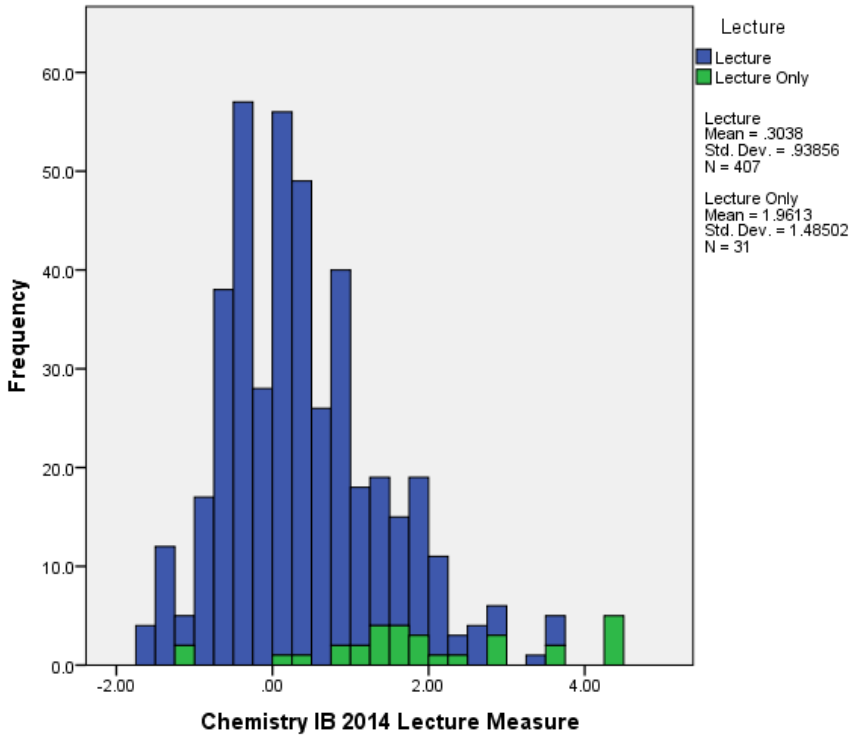

Figure 597: The Student Ability of Students on Overlapping Items within Chemistry IB Lecture Test MCQ Assessments in 2014 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

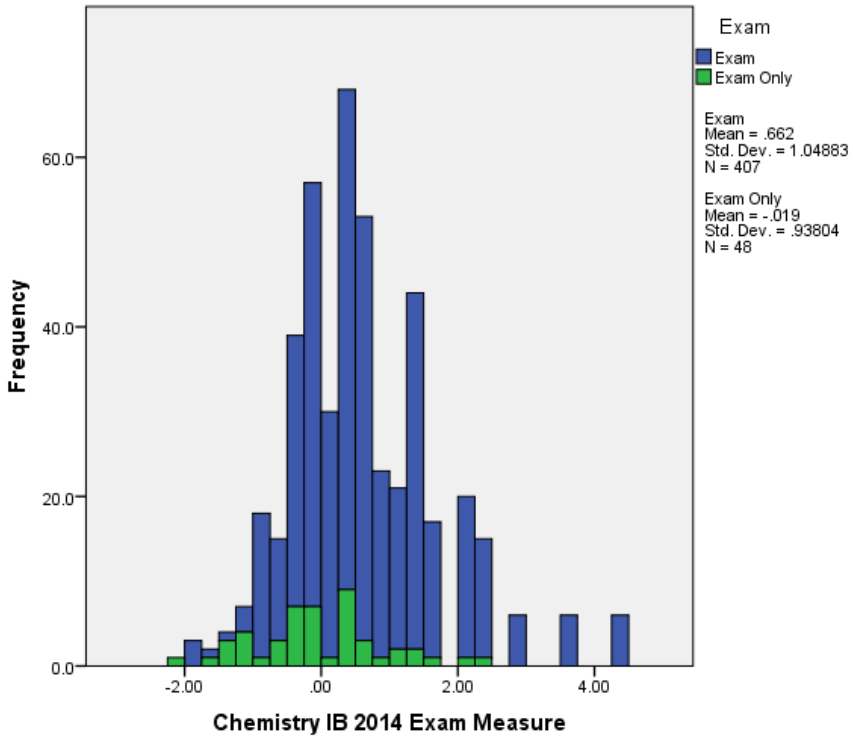

Figure 598: The Student Ability of Students on Overlapping Items within Chemistry IB Redeemable Exam MCQ Assessments in 2014 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

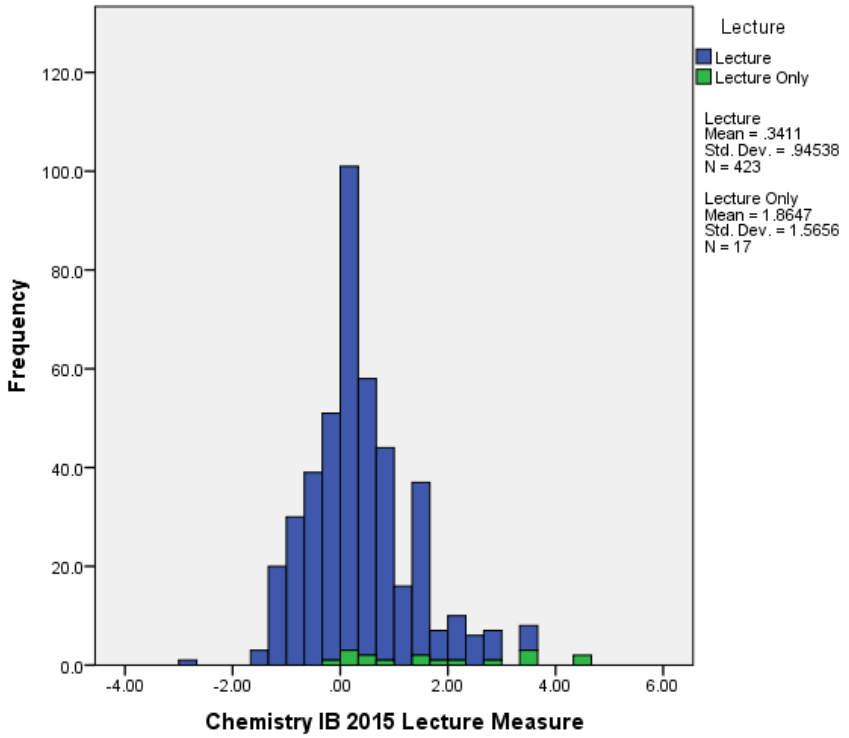

Figure 599: The Student Ability of Students on Overlapping Items within Chemistry IB Lecture Test MCQ Assessments in 2015 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

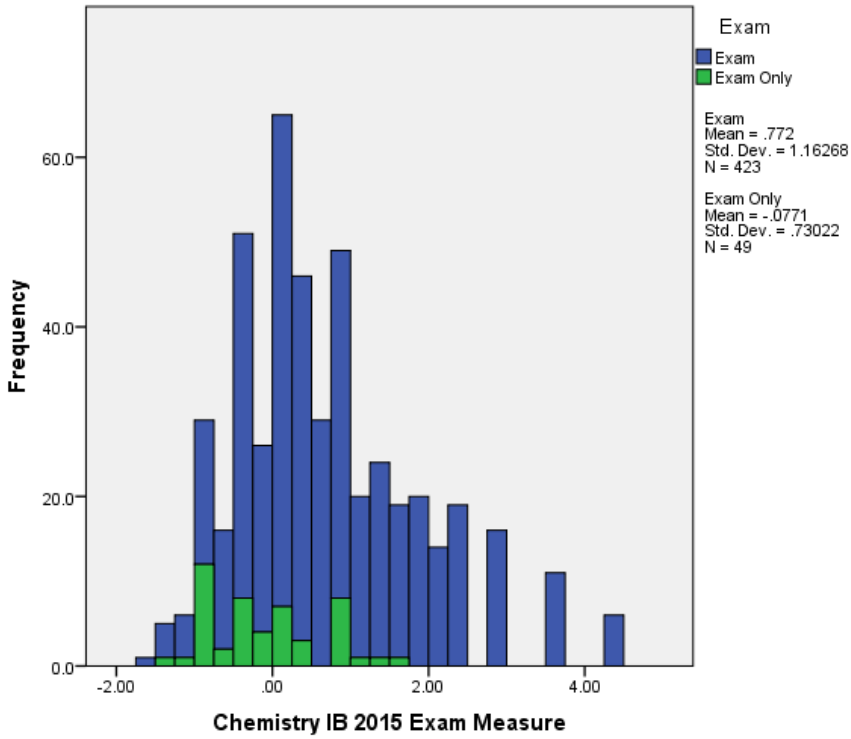

Figure 600: The Student Ability of Students on Overlapping Items within Chemistry IB Redeemable Exam MCQ Assessments in 2015 Comparing Students Who Only Undertook One Assessments to those Who Undertook Both

7.25 Comparison of Shared Items Difficulty using CTT Over Multiple Years

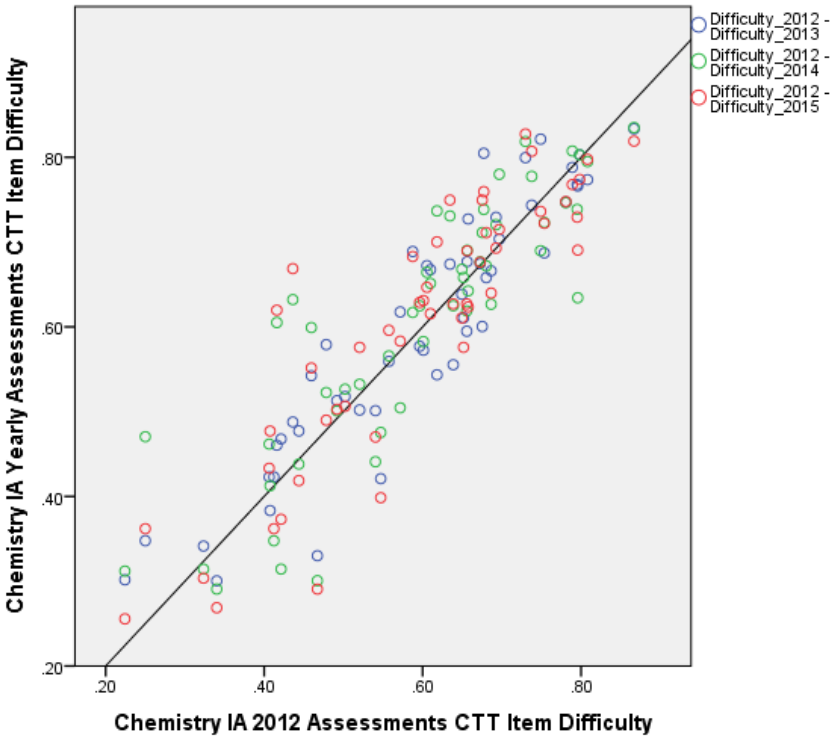

Figure 601: Cross-Plot of CTT Item Difficulty Using Overlapping Items from all Years of Chemistry IA Analysed

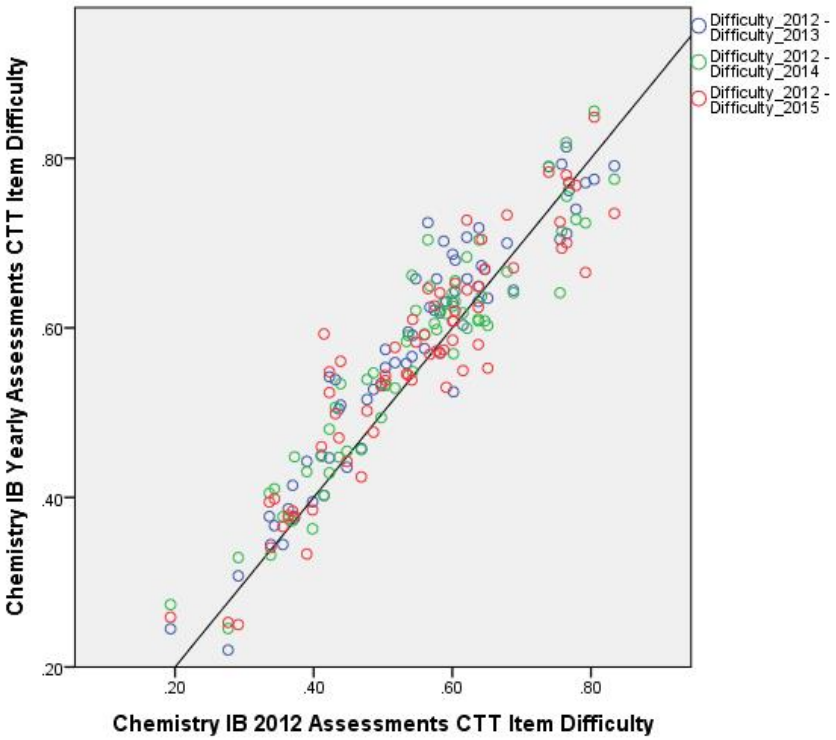

Figure 602: Cross-Plot of CTT Item Difficulty Using Overlapping Items from all Years of Chemistry IB Analysed

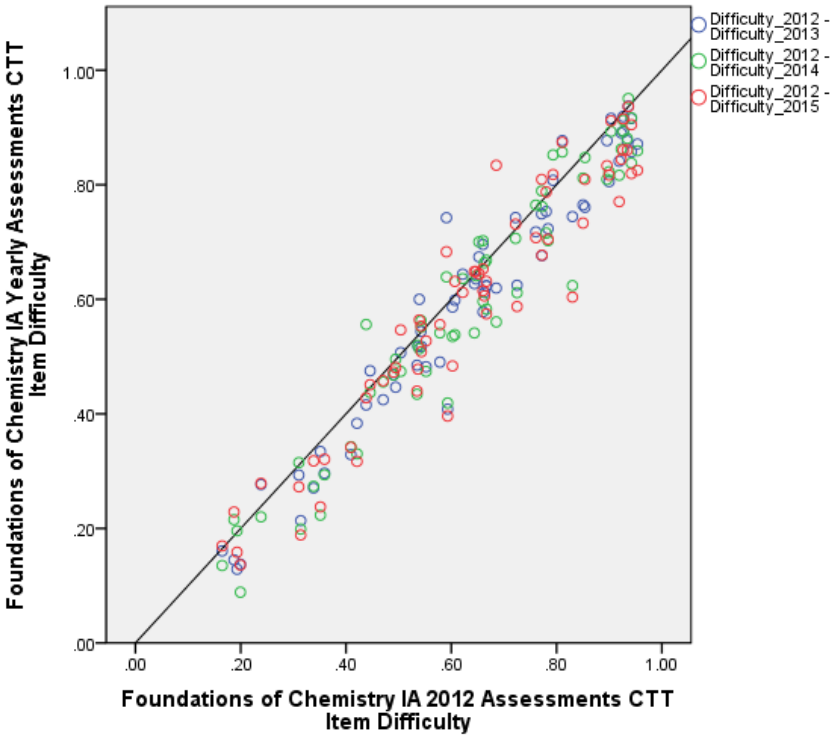

Figure 603: Cross-Plot of CTT Item Difficulty Using Overlapping Items from all Years of Foundations of Chemistry IA Analysed

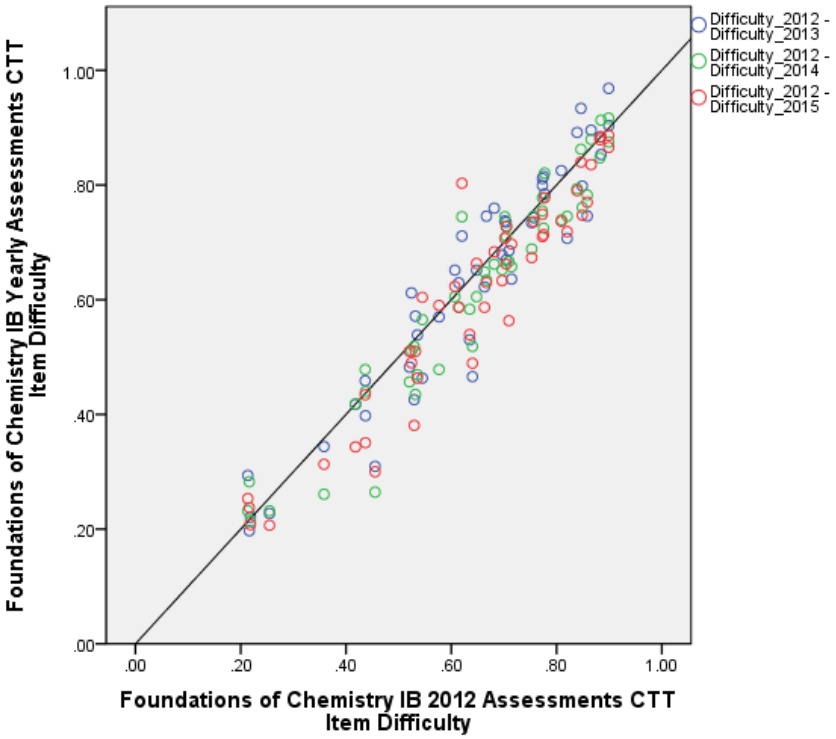

Figure 604: Cross-Plot of CTT Item Difficulty Using Overlapping Items from all Years of Foundations of Chemistry IB Analysed

7.26 Comparison of Student Percentage Score Distribution in MCQ Assessments Over Multiple Years

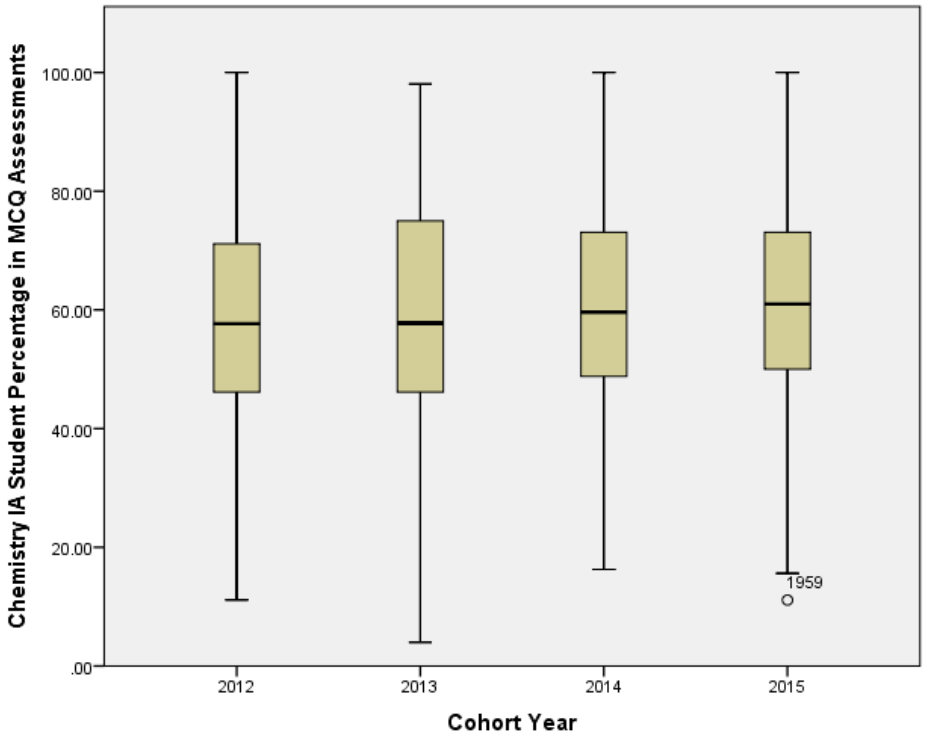

Figure 605: Boxplot of Student Percentage Distribution from Overlapping MCQ Assessment Items within Chemistry IA Comparing all years Analysed

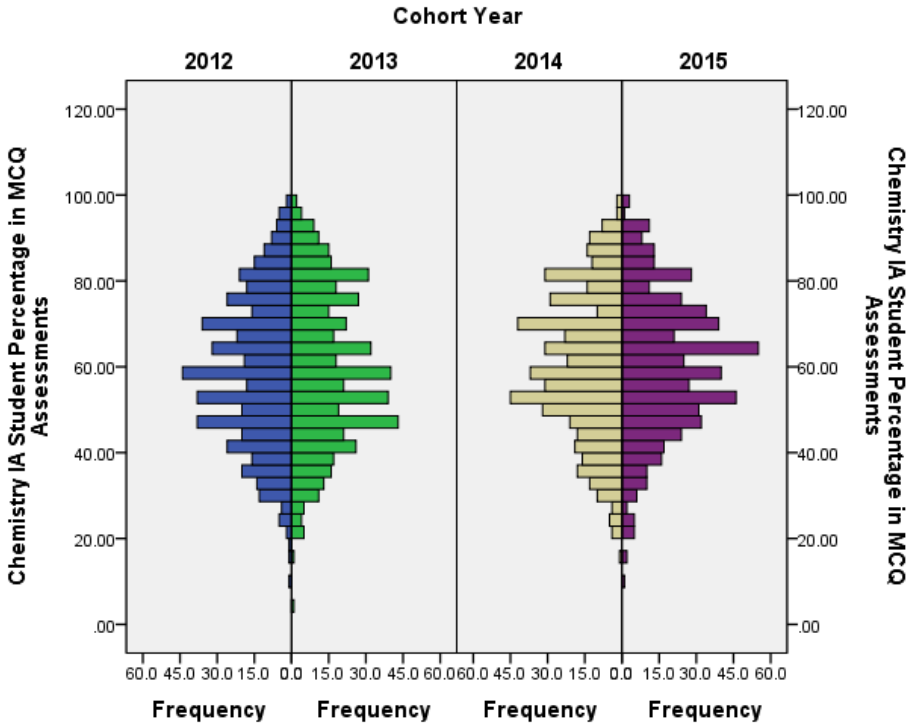

Figure 606: Histogram Distribution of Student Percentage Score from Overlapping MCQ Assessment Items within Chemistry IA Comparing all years Analysed

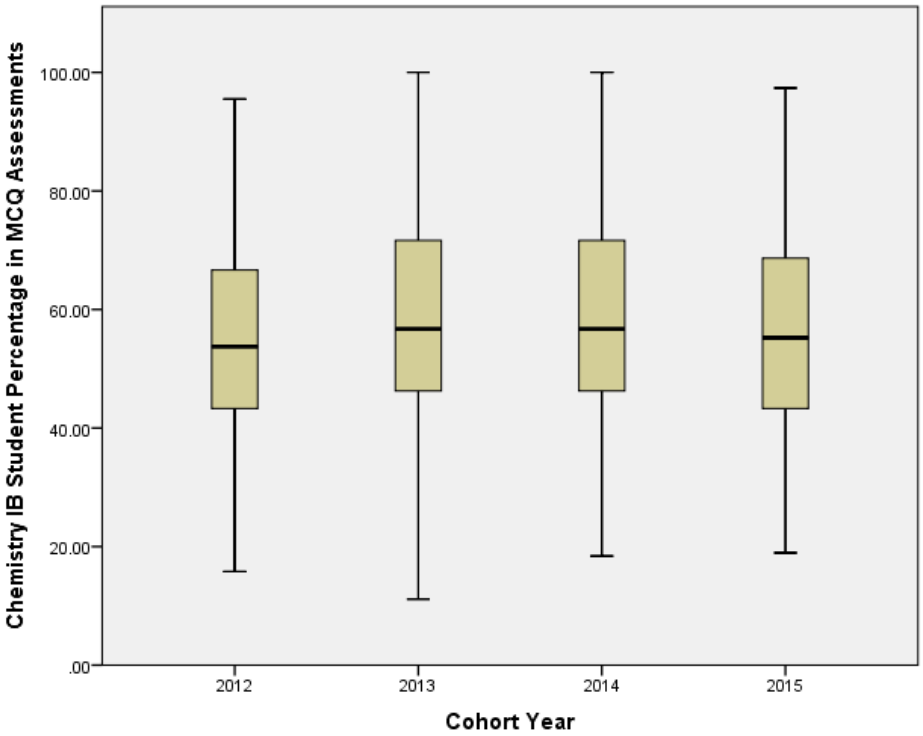

Figure 607: Boxplot of Student Percentage Distribution from Overlapping MCQ Assessment Items within Chemistry IB Comparing all years Analysed

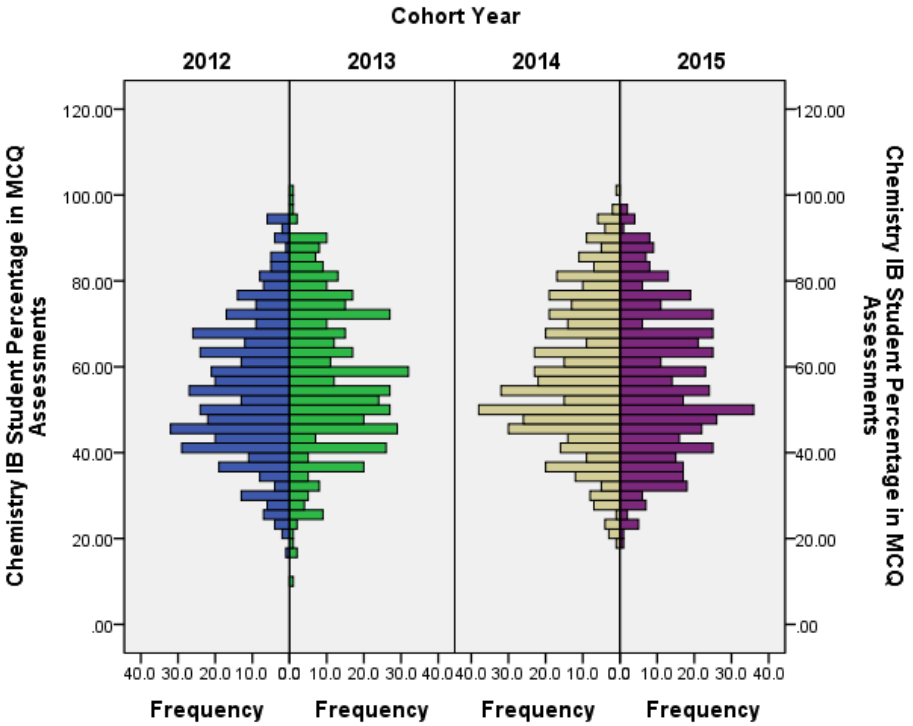

Figure 608: Histogram Distribution of Student Percentage Score from Overlapping MCQ Assessment Items within Chemistry IB Comparing all years Analysed

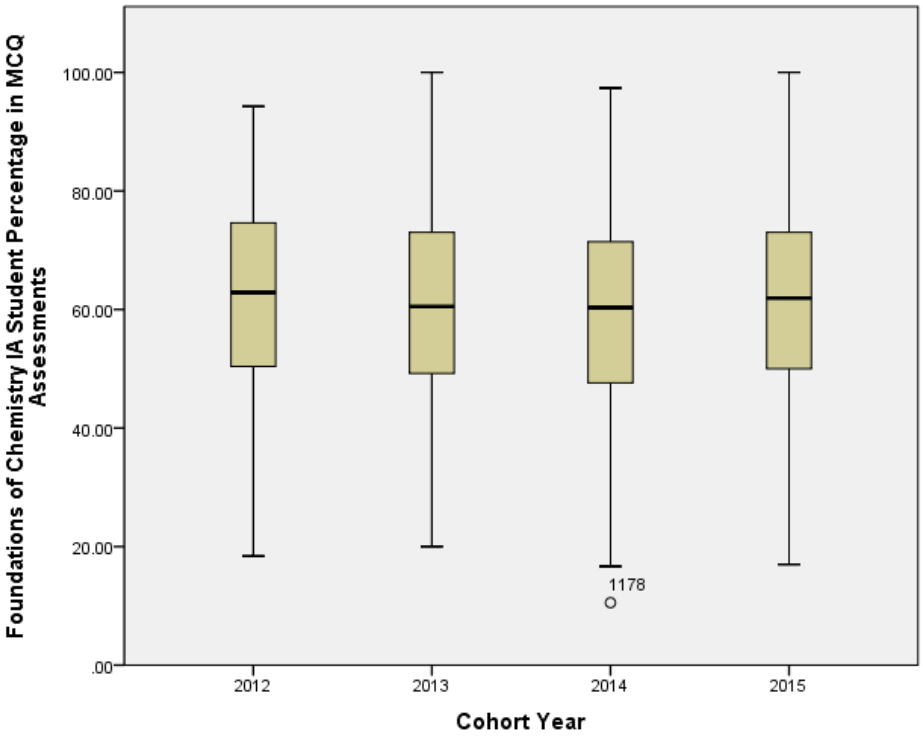

Figure 609: Boxplot of Student Percentage Distribution from Overlapping MCQ Assessment Items within Foundations of Chemistry IA Comparing all years Analysed

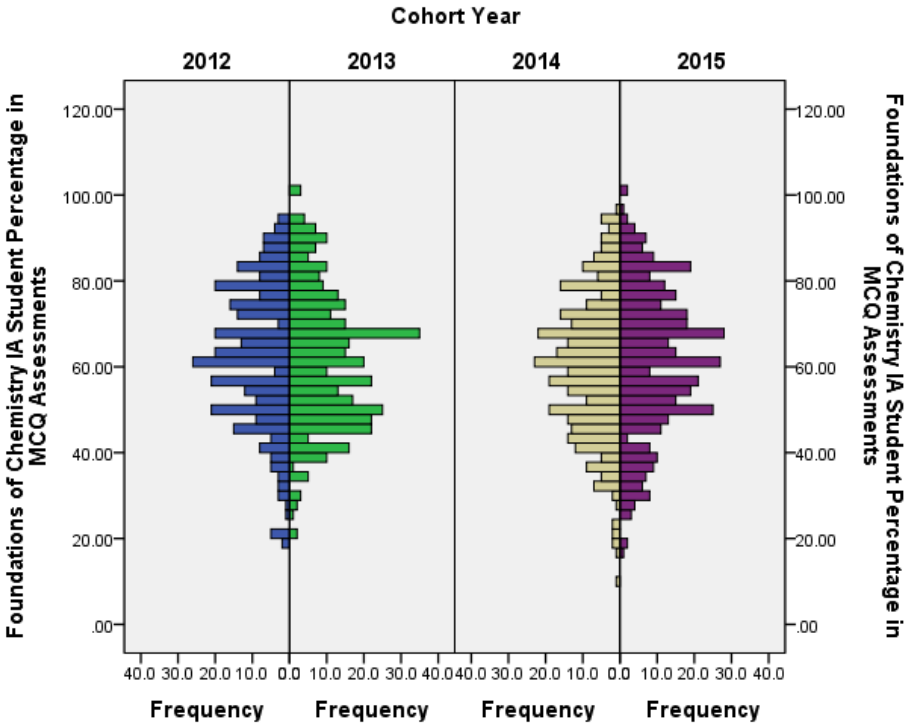

Figure 610: Histogram Distribution of Student Percentage Score from Overlapping MCQ Assessment Items within Foundations of Chemistry IA Comparing all years Analysed

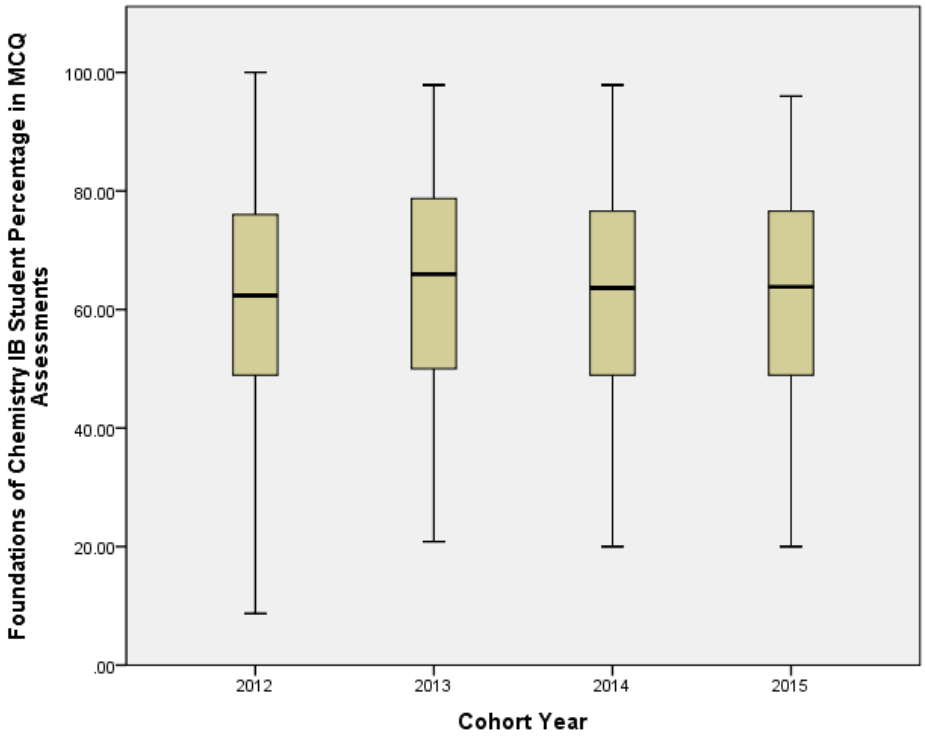

Figure 611: Boxplot of Student Percentage Distribution from Overlapping MCQ Assessment Items within Foundations of Chemistry IB Comparing all years Analysed

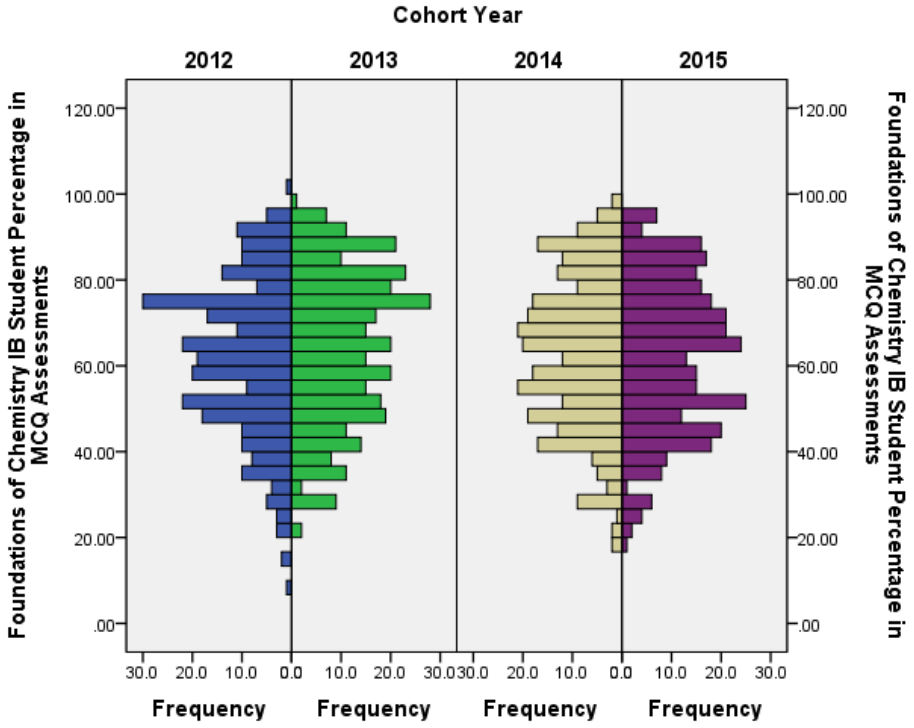

Figure 612: Histogram Distribution of Student Percentage Score from Overlapping MCQ Assessment Items within Foundations of Chemistry IB Comparing all years Analysed

7.27 Comparison of Shared Items Difficulty using Rasch Analysis Over Multiple Years

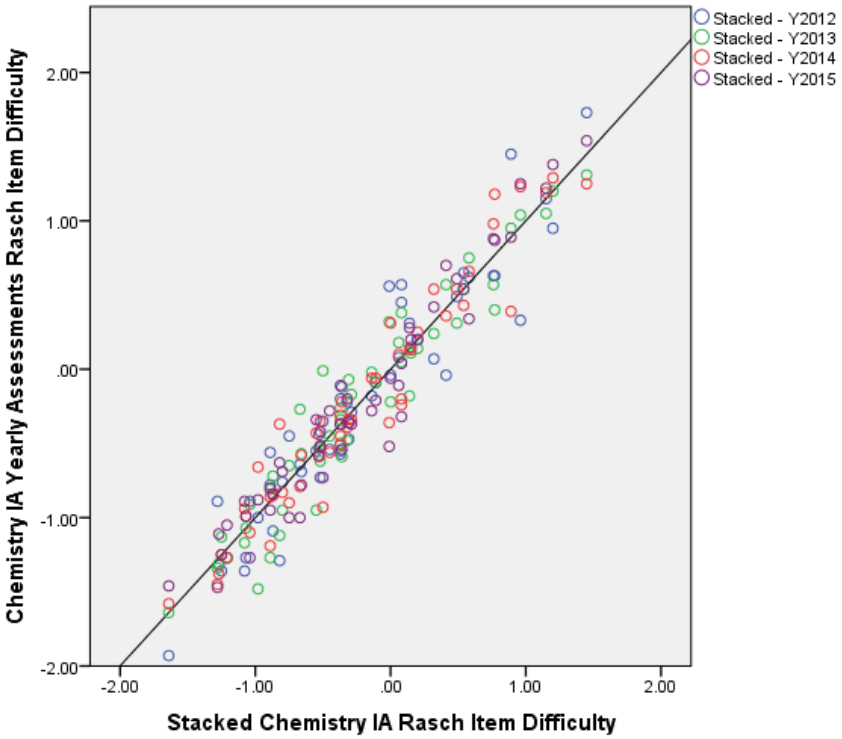

Figure 613: Cross-Plot of Rasch Analysis Item Difficulty Measure Using Overlapping Items from all Years of Chemistry IA Analysed

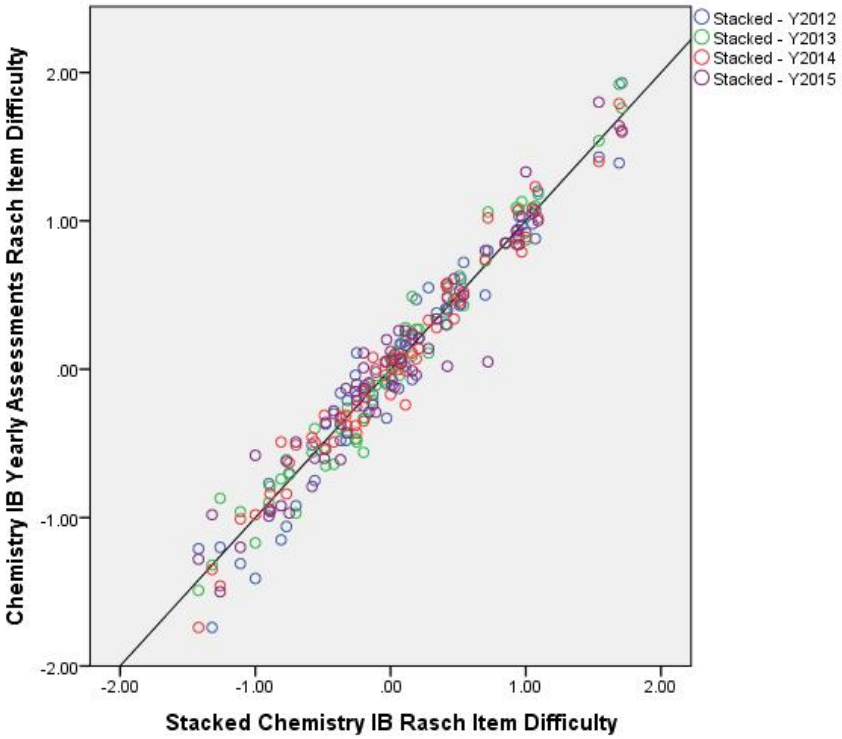

Figure 614: Cross-Plot of Rasch Analysis Item Difficulty Measure Using Overlapping Items from all Years of Chemistry IB Analysed

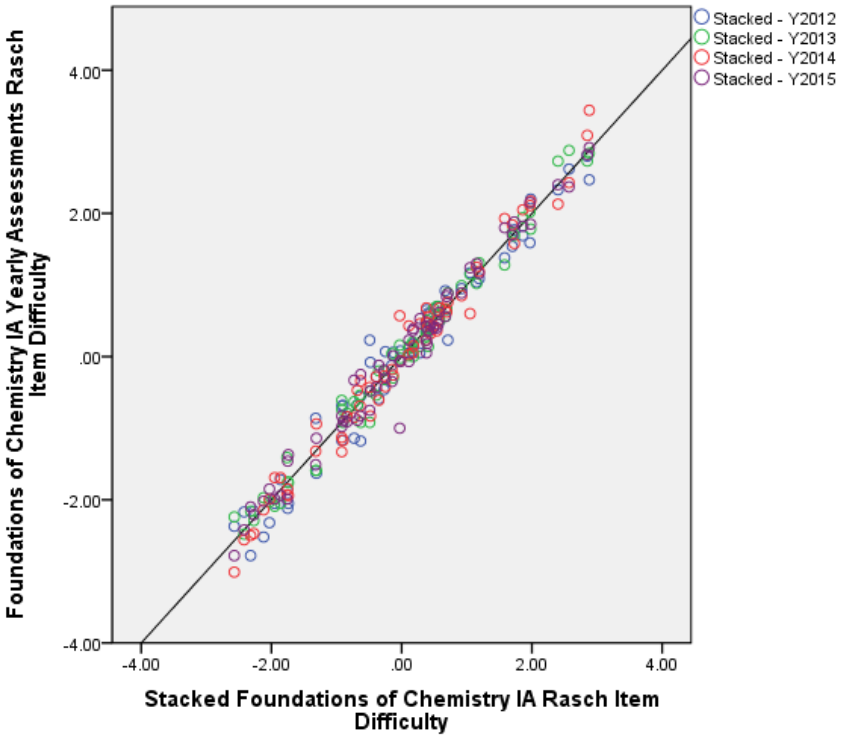

Figure 615: Cross-Plot of Rasch Analysis Item Difficulty Measure Using Overlapping Items from all Years of Foundations of Chemistry IA Analysed

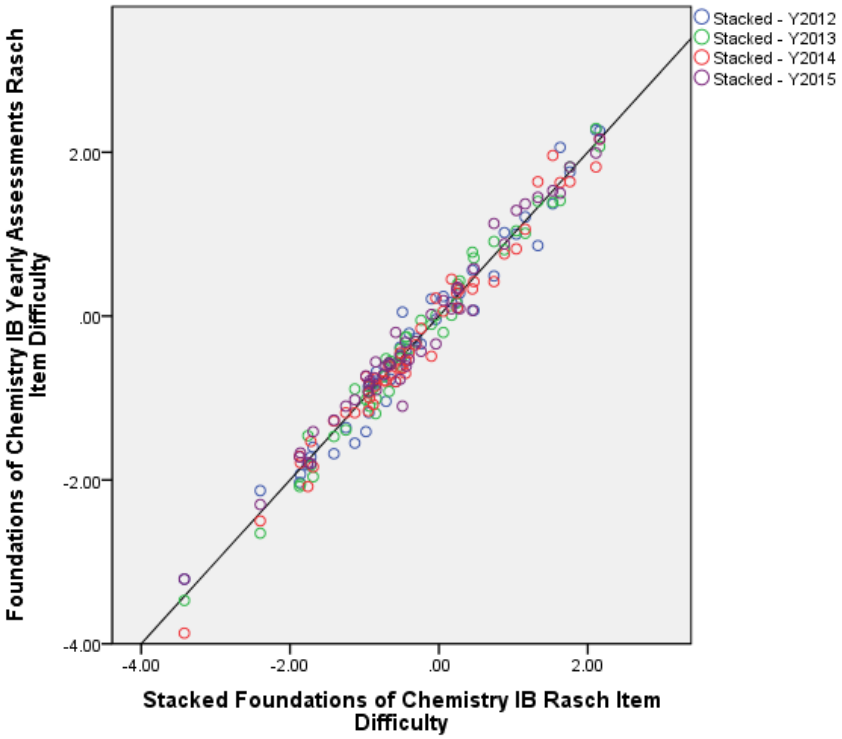

Figure 616: Cross-Plot of Rasch Analysis Item Difficulty Measure Using Overlapping Items from all Years of Foundations of Chemistry IB Analysed

7.28 Comparison of Student Rasch Ability Measure Distribution Over Multiple Years

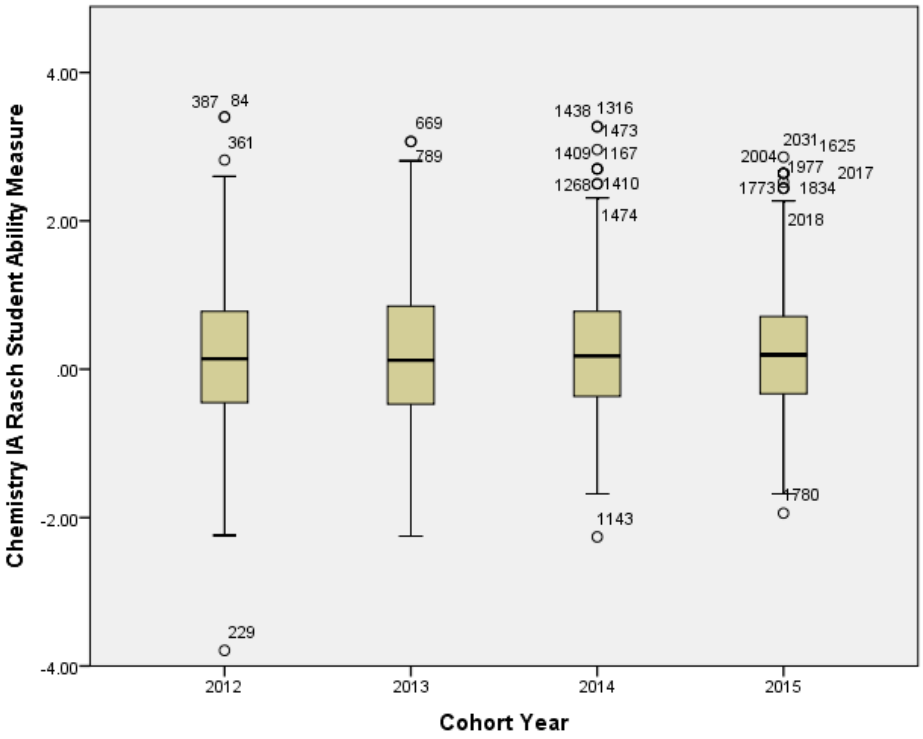

Figure 617: Boxplot of Student Ability Measure Distribution from MCQ Assessment within Chemistry IA Comparing all years Analysed

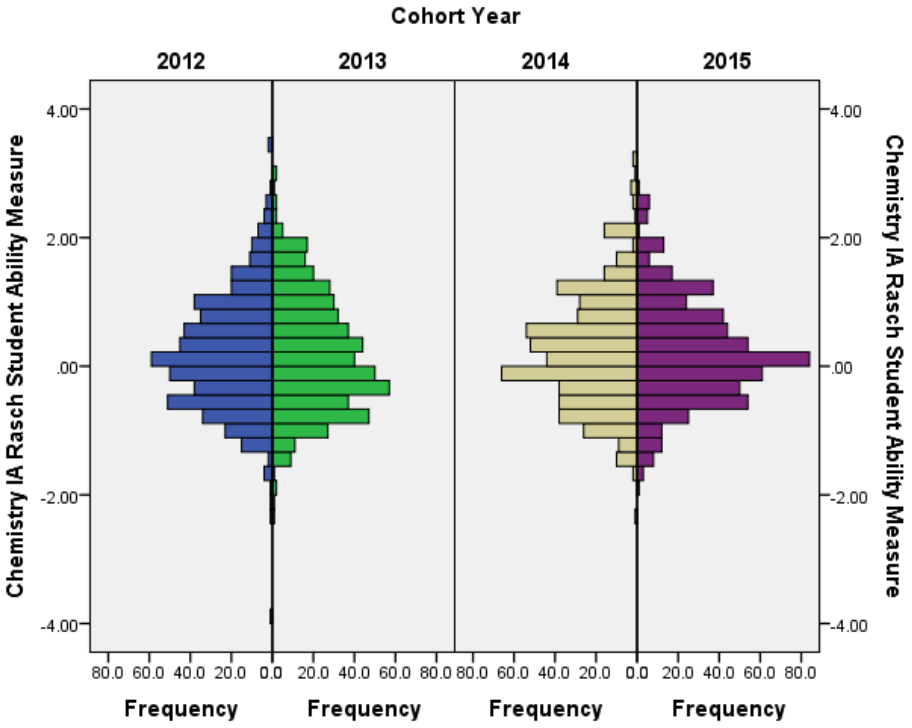

Figure 618: Histogram Distribution of Student Ability Measure from MCQ Assessment within Chemistry IA Comparing all years Analysed

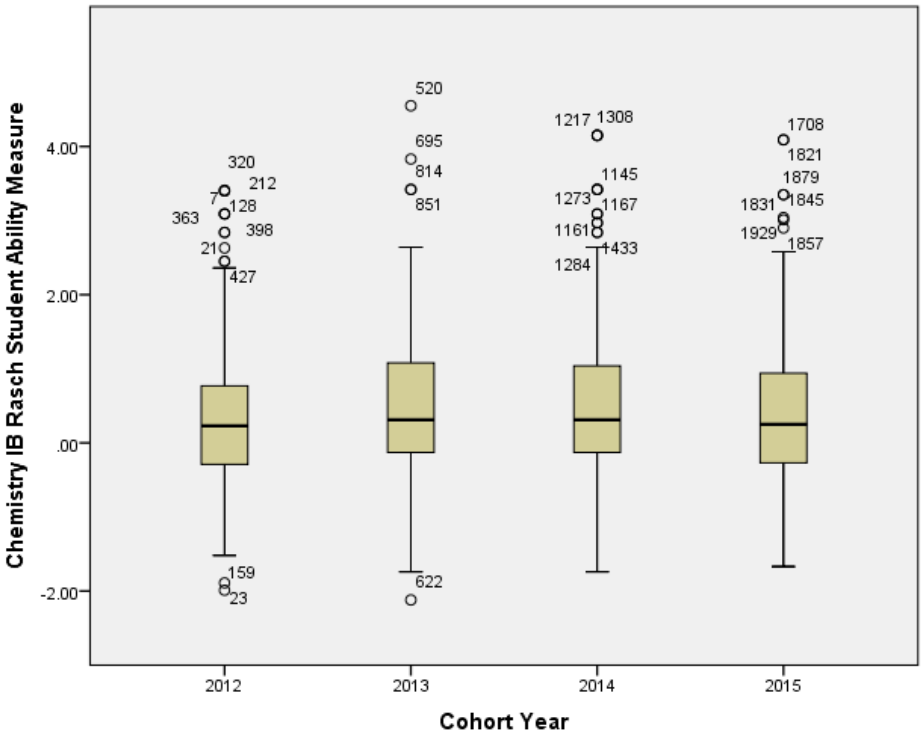

Figure 619: Boxplot of Student Ability Measure Distribution from MCQ Assessment within Chemistry IB Comparing all years Analysed

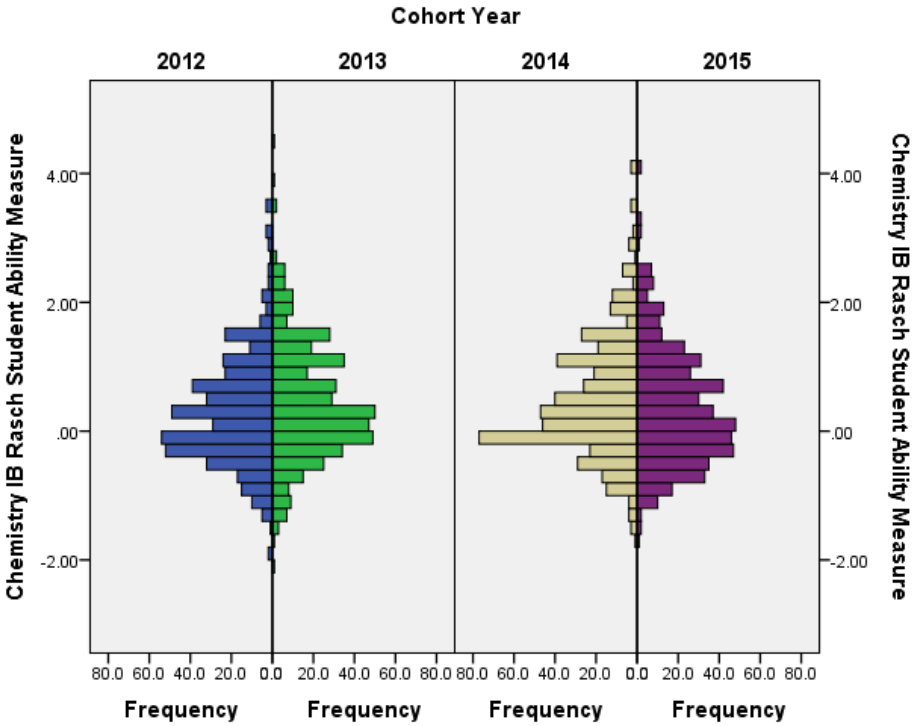

Figure 620: Histogram Distribution of Student Ability Measure from MCQ Assessment within Chemistry IB Comparing all years Analysed

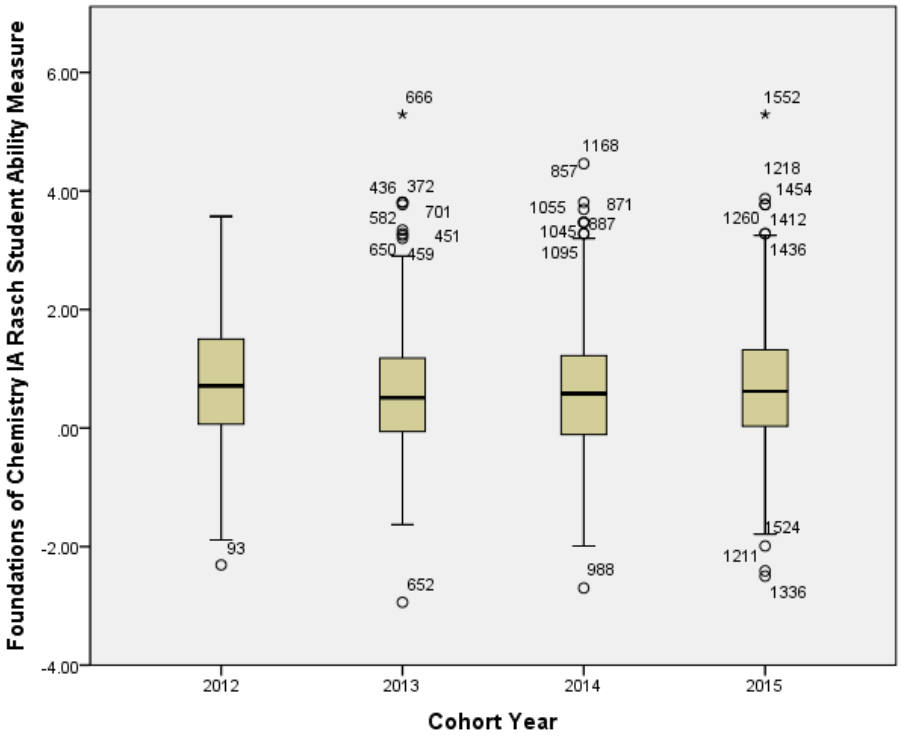

Figure 621: Boxplot of Student Ability Measure Distribution from MCQ Assessment within Foundations of Chemistry IA Comparing all years Analysed

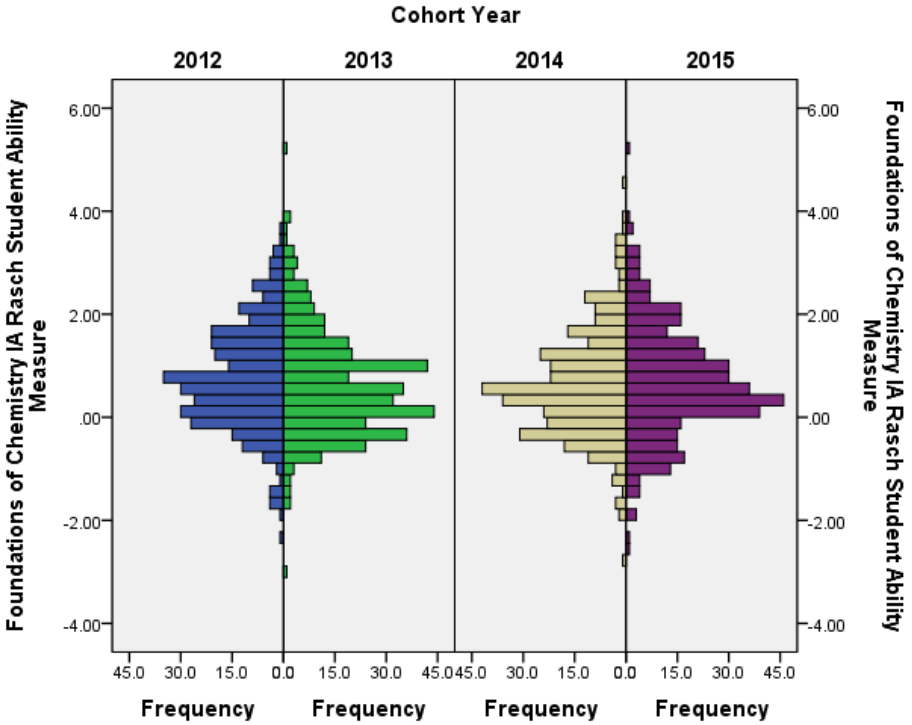

Figure 622: Histogram Distribution of Student Ability Measure from MCQ Assessment within Foundations of Chemistry IA Comparing all years Analysed

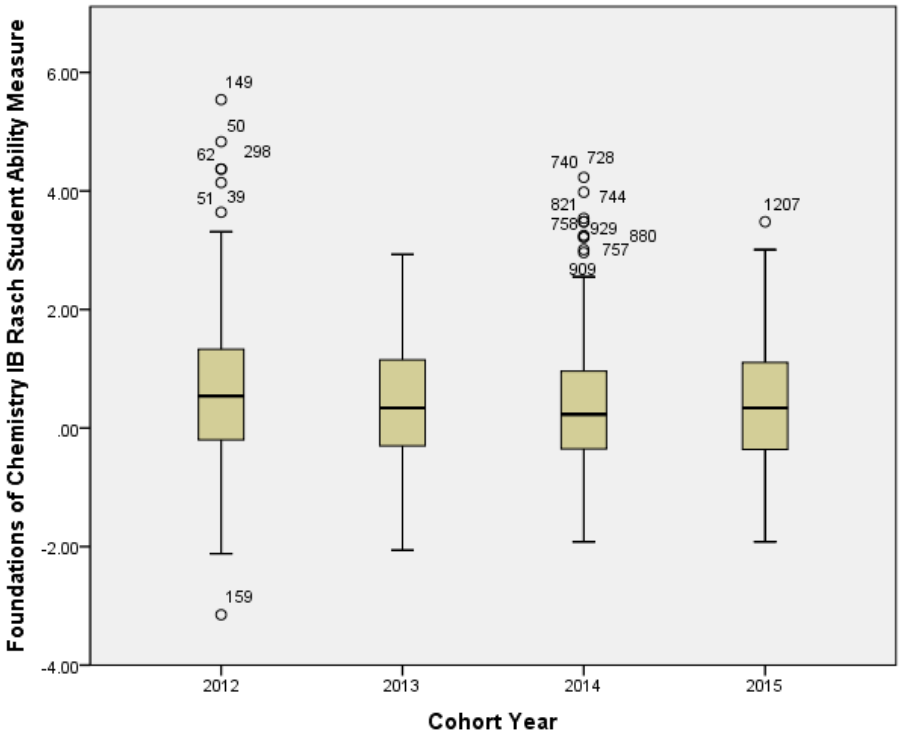

Figure 623: Boxplot of Student Ability Measure Distribution from MCQ Assessment within Foundations of Chemistry IB Comparing all years Analysed

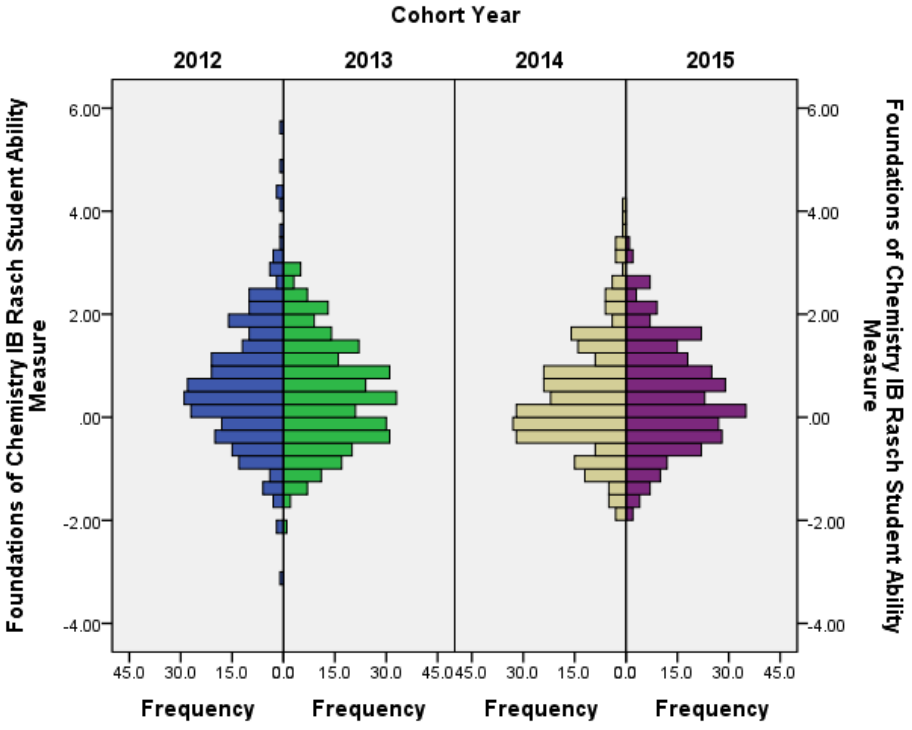

Figure 624: Histogram Distribution of Student Ability Measure from MCQ Assessment within Foundations of Chemistry IB Comparing all years Analysed